Ф. П ТАРАСЕНКО

НЕПАРАМЕТРИЧЕСКАЯ СТАТИСТИКА

ИЗДАТЕЛЬСТВО ТОМСКОГО УНИВЕРСИТЕТА
Томск — 1976

Многие достижения современной физики, радиотехники, технической кибернетики, измерительной техники, дефектоскопии, геофизики, биологии, медицины, социологии и других наук, связанных с обработкой экспериментальных данных, основаны на применении разнообразных статистических методов. Сейчас все большую актуальность приобретают статистические задачи при высокой априорной неопределенности, когда практически ничего неизвестно о виде функций, распределения величин, участвующих в задаче. Потребности в решении таких задач отвечает непараметрическая статистика. В данной книге сделана попытка изложить все основные разделы этой ветви математической статистики, а некоторые ее главы посвящены вопросам, ранее не освещавшимся в монографической литературе. Книга будет полезной всем имеющим дело со статистикой и желающим пользоваться ею осознанно.

2-2

ОГЛАВЛЕНИЕ

Глава X. **О ранговых тестах**

ПРЕДИСЛОВИЕ

Последние десятилетия развития прикладных теоретических и экспериментальных наук характеризуются непрерывным возрастанием полноты и строгости учета случайных факторов. Многие достижения современной физики, радиотехники, биологии, медицины, технической кибернетики, социологии, измерительной техники, геофизики, дефектоскопии и других наук, связанных с обработкой экспериментальных данных, основаны на применении разнообразных статистических методов. Поставщиком этих методов является математическая статистика.

С точки зрения потребителей статистических методов, математическую статистику можно представить как некий «черный ящик», т. е. как систему, внутри которой неважно, что и как делается, а важно то, что на ее выходе имеются различные статистические процедуры и инструкции по пользованию этими процедурами. Характерным становится то, что практика все чаще выдвигает задачи, которые не укладываются в требования этих инструкций. Так возникает «сигнал рассогласования», требующий развития самой математической статистики.

Главной причиной необходимости создания новых статистических методов является расхождение между моделью, на которой базируются имеющиеся методы, и реальной ситуацией. Значительная часть результатов математической статистики основана на предположении о том, что информации, имеющейся у потребителя, достаточно для представления участвующих в задаче распределений в виде некоторых функций с конечным числом параметров. Однако на практике это предположение часто оказывается нереальным. Наиболее ярко такая ситуация проявляется в социологии и биологии, но в последние годы даже представители техни-

ческих дисциплин стали (с некоторым удивлением) обнаруживать, что это иногда относится и к ним. Наглядный пример тому дает история исследований распространения радиоволн по реальным трассам с рассеянием на неоднородностях тропосферы и ионосферы.

Потребности в создании статистических процедур, не предполагающих знания вида распределений, и отвечает ветвь математической статистики, получившая название непараметрической статистики Непараметрическая статистика прямо, открыто, с самого начала, в самой исходной модели рассматривает только такие ситуации, в которых о функциональном виде распределений ничего не известно Единственной доступной потребителю априорной информацией считается информация о характере случайной величины (например, непрерывна она или дискретна) и о типе различия между распределениями. Сами распределения могут быть какими угодно, а нас интересует лишь, сдвинуты ли они относительно друг друга; различаются ли они масштабами, наличием или отсутствием симметрии или зависимости; наконец, есть ли между ними вообще какое бы то ни было отличие, или они могут считаться одинаковыми — приблизительно таков круг задач, решаемых непараметрической статистикой

Интерес практиков и теоретиков к непараметрической статистике подогревается не только тем, что она исходит из более широкой и реалистической модели, нежели классические ветви статистики; не только тем, что получаемые ею процедуры часто очень просты в реализации; не только тем, что в ней трудные математические проблемы решаются изящно и красиво. Неожиданным, удивительным и чрезвычайно стимулирующим оказалось то, что несмотря на очень малый объем априорной информации, используемой при построении непараметрических процедур, они обладают высокой эффективностью: 1) потери эффективности при переходе от параметрических к непараметрическим процедурам (в случае истинности параметрической модели!) незначительны и довольно часто составляют всего несколько процентов; 2) эффективность непараметрической процедуры по сравнению с фиксированной параметрической резко возрастает при отклонении истинных распределений от расчетных; 3) во многих случаях непараметрические процедуры оказываются асимптотически оптимальными. Главным источником этих замечательных свойств является то, что при непараметрическом подходе используется только достоверная, доступная нам априорная информация Недостоверная информация (например, предположение о нормальности распределения, тогда как оно на самом деле негауссово) в некотором смысле эквивалентна введению дополнительных случайных, мешающих факторов.

Эти качества непараметрических процедур объясняют возрастающий интерес к непараметрической статистике и, несомненно, обеспечат широкое внедрение ее результатов в практику в ближайшем будущем. Непараметрическая статистика бурно развивается: число журнальных статей в этой области насчитывает уже тысячи; за рубежом появились и первые, пока немногочисленные, монографии по различным разделам непараметрической статистики (одна из них переведена на русский язык*). В данной книге делается попытка изложить все основные разделы непараметрической статистики с единых методических позиций

Структура книги такова: Часть I посвящена обсуждению общих вопросов непараметрической статистики (ее отличию от остальных ветвей математической статистики; изложению методологии непараметрической статистики; сюда также включено для удобства читателя описание способов сравнения статистических процедур); в Части II излагаются те статистические свойства выборок и их преобразований, на которых основаны непараметрические методы; Часть III содержит описание непараметрических процедур и их свойств Автор стремился по возможности полно осветить современное состояние непараметрической статистики: в книгу вошли многие результаты различных авторов; однако список литературы не претендует на полноту. Некоторые параграфы книги содержат результаты самого автора.

Книга рассчитана на широкий круг научных и технических специалистов тех профилей, которым приходится иметь дело со статистикой. Чтение книги предполагает знание теории вероятностей и общее знакомство с математической статистикой. Математический уровень книги можно назвать «промежуточным»: противоречие между желанием расширить круг потенциальных читателей книги и желанием осветить важные вопросы достаточно глубоко привело к целесообразности остановиться на том уровне строгости, который обычно принят, например, в физике (хотя надо сказать, что в любой области требования к математической строгости непрерывно повышаются).

Несколько слов об оформлении книги. Нумерация параграфов и формул такова: § 9 11 означает одиннадцатый параграф девятой главы, (9.11.5) означает пятую формулу параграфа 9 11. Аналогично табл 9.11 2 означает вторую таблицу параграфа 9.11. Список литературы составлен по алфавитному списку авторов, нумерация ведется отдельно для

---

* Это монография Я Гаека и З Шидака «Теория ранговых критериев». Изд во «Наука», М., 1971

каждого автора. Например, ссылка «Вайсс [3]» означает статью, приведенную под номером 3 в списке работ Вайсса.

В заключении Предисловия автор считает своим приятным долгом выразить благодарность и признательность всем лицам, так или иначе способствовавшим появлению данной книги. Это, во-первых, руководство факультета прикладной математики и радиофизического факультета Томского университета, согласившееся на постановку спецкурсов по непараметрической статистике: книга, по существу, есть углубленный и расширенный вариант курса лекций, который автор неоднократно читал в течение трех лет. Далее, мне хотелось бы с благодарностью упомянуть коллег по профессии, беседы с которыми помогли мне улучшить многие разделы книги: москвичей М. С. Пинскера, Б. Р. Левина, К. Ш. Зигангирова, Р. Л. Хасьминского, Д. М. Чибисова, Л. Ф. Бородина; Ю. Н. Линькова (Донецк), Л. Д. Дэвиссона (США), А. Переза (Чехословакия); полезной была также переписка с Я. Гаеком (Чехословакия) и Ю. Куликовским (Польша). В работе над книгой мне помогали мои сотрудники по Сибирскому физико-техническому институту (СФТИ) при ТГУ. Эта помощь выражалась в многочисленных дискуссиях по отдельным главам и параграфам, в совместном выполнении некоторых исследований, результаты которых вошли в книгу. Здесь я должен упомянуть Ю. Г. Дмитриева, А. П. Серых, Г. М. Кошкина, В. А. Симахина. Особо хочу поблагодарить В. П. Шуленина, с которым мы вместе работали над многими параграфами

Д. М. Чибисов и Б. Р. Левин прочли рукопись книги и сделали много критических замечаний, за которые автор им весьма благодарен. Однако и после некоторых изменений книга не свободна от недостатков, за которые автор несет ответственность единолично и с благодарностью примет критику в свой адрес.

Ф Тарасенко

Часть I

ОБЩИЕ ВОПРОСЫ НЕПАРАМЕТРИЧЕСКОЙ СТАТИСТИКИ

ГЛАВА I

ОСНОВНЫЕ ПОНЯТИЯ И МЕТОДЫ НЕПАРАМЕТРИЧЕСКОЙ СТАТИСТИКИ

§ 1.1. ОБЩАЯ СХЕМА ПРИНЯТИЯ СТАТИСТИЧЕСКИХ РЕШЕНИЙ

Истина редко доступна человеку в явном виде. Обычно о ней приходится судить по некоторым связанным с нею косвенным данным, на которые оказывает влияние не только то, что нас интересует, но и посторонние, «мешающие» факторы.

В науке и технике наблюдаемые данные обычно носят вполне конкретный характер: это либо точная констатация факта (типа «в результате опыта данное событие произошло (или — не произошло)»), либо определенное число, являющееся результатом единичного эксперимента. По результатам последовательности таких экспериментов необходимо вынести суждение об интересующей нас и недоступной прямому наблюдению ситуации. Математическая статистика является теорией, указывающей, как нужно строить процедуры вынесения решений в условиях стохастической неопределенности (синтез) и какими свойствами будут обладать получаемые решения (анализ).

Для того, чтобы охарактеризовать структуру математической статистики в целом и место непараметрической статистики в ней, рассмотрим общую схему принятия статистических решений, которая представлена на рис. 1.1.1.

Введем некоторый абстрактный параметр $\theta \in \Theta$, характеризующий состояние изучаемого объекта. Абстрактность параметра θ состоит в том, что он в одних случаях может мыслиться как обычный числовой (одно- или многомерный) параметр, в других — просто как индекс некоторой ситуации, не обязательно характеризуемой количественно. Пространство Θ всевозможных θ назовем пространством ситуаций; истинное «значение» θ — это именно то, что мы хотели бы знать.

При недоступности непосредственного наблюдения θ нам приходится наблюдать реализовавшиеся в опыте значения $x_1, x_2, \ldots, x_N$ некоторой случайной величины X, статистически

связанной с Θ Связь X с Θ обеспечивает возможность извлечения из выборки $\beta = (x_1, x_2, \ldots, x_N)$ информации о θ, а случайность X вызвана влиянием посторонних неконтролируемых факторов. Такой переход от θ к β можно изобразить как результат некоторой операции $\mu(\theta, n)$ над θ и случайным процессом n, генерируемым эквивалентным источником стохастичности (см. рис. 1.1.1). Множество B всех β, которые могут реализоваться при всевозможных θ и n, назовем пространством наблюдений.

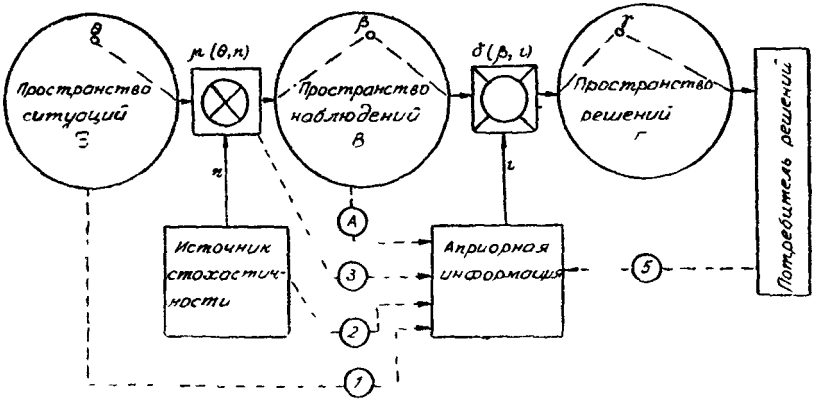

Рис. 1 1 1

Информация о θ извлекается из β с помощью некоторой статистической процедуры, или решающей функции δ, и выдается в виде решения γ о том, каково же «значение» θ. Множество всех решений γ (т. е. пространство решений Г) включает в себя все элементы пространства Θ, а иногда и дополнительные решения (например, «данных недостаточно для суждения о θ с заданной надежностью», «необходима рандомизация» и т. п.).

Оператор δ выполняет функцию обращения оператора μ по отношению к θ. Насколько такое обращение удается осуществить, в сильной степени зависит от априорной информации i, которая используется решающей функцией δ наряду с информацией из эксперимента β. Идеальной была бы такая процедура, при которой обеспечивается тождество $\gamma \equiv \theta$ (или, в операторном описании, $\delta\mu \equiv 1$). Однако оказывается, что это возможно лишь в редких (как правило, тривиальных) случаях; обычно же γ является случайной величиной (или функцией, или событием), связанной с θ. Разные процедуры обеспечивают разную «близость» γ к θ, т. е. дают решения различного качества. Естественно желание отыскать

наилучшие процедуры, для этого приходится вводить количественные меры соответствия пространств Θ и Γ и разрабатывать методы синтеза оптимальных δ. И тут вновь главную роль играет априорная информация о решаемой статистической задаче. Рассмотрим подробнее роль априорной информации в синтезе статистических процедур и влияние априорной информации на их качество

§ 1.2. УРОВНИ АПРИОРНОЙ ИНФОРМАЦИИ И РАЗЛИЧНЫЕ ВЕТВИ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Известно, что без априорной информации нельзя извлечь апостериорную. Это фундаментальное положение лишь на первый взгляд кажется очевидным. При его ближайшем рассмотрении возникает ряд непростых проблем как естественнонаучного, так и философского характера. Например: если в качестве априорных сведений для данного эксперимента берутся апостериорные сведения предыдущего опыта, то что служит априорной информацией для самого первого опыта? Как вообще появляются новые, неизвестные ранее знания? Иного типа вопросы возникают в связи со следующим специфическим соотношением между априорной и апостериорной информацией: если априорной информации слишком много, эксперимент вообще не нужен (все и без него известно), а если ее слишком мало, то из эксперимента либо вообще невозможно извлечь полезную информацию, либо доступной оказывается ее очень скудная доля. Чем в таком случае определяется оптимальный уровень априорной информации, обеспечивающий максимальный информационный «урожай», снимаемый с эксперимента? И т. д. и т. п. Интересующихся этими и подобными вопросами отошлем к общенаучным (например, Л. Бриллюэн [1], [2]), философским (С. Л. Соболев, А И Китов, А. А. Ляпунов [1]; Ф. П. Тарасенко [1] и др.) работам или к работам по планированию экспериментов.

Определим конкретнее, что именно понимается под априорной информацией в статистике. Обратимся к общей схеме принятия статистических решений (см. рис. 1.1.1). В качестве априорной информации выступают *любые* сведения, имеющиеся у статистика *до того*, как фиксируются результаты эксперимента β и *до того*, как он приступит к синтезу (или выбору) процедуры δ. Эти сведения характеризуют: 1) пространство ситуаций Θ; 2) природу случайных факторов n; 3) оператор μ, т. е. тип взаимодействия мешающих факторов n с параметром ситуации θ; 4) пространство наблюдений B. На рис. 1.1.1 соответствующие информационные каналы изображены пунктирными линиями под теми же

номерами. Кроме того, на рис. 1.1.1 изображен и канал, по которому поступает априорная информация от потребителя решений. Обсудим сначала именно эту часть схемы.

Статистика, конечно, интересуют и чисто «внутренние» задачи теории, направленные на ее собственное развитие. Но поскольку конечной, «внешней» целью этого развития является предложение процедур для решения практических задач, теория должна учитывать и интересы потребителя. Всякое решение выносится для того, чтобы его в дальнейшем как-то использовать. Как именно — это зависит от потребителя, и предугадать все возможные случаи немыслимо. Для учета требований потребителя в теории предусмотрена возможность ввода информации от него. Эта информация может быть задана в виде функции потерь, которые понесет потребитель при неправильных решениях; в виде подходящего ему критерия оптимальности процедуры; в виде ограничений, налагаемых его возможностями на реализацию процедур, и т. п. Понятно, что эта априорная информация существенно влияет на выбор методов и результаты синтеза решающей функции δ.

Обратимся теперь к рассмотрению статистической априорной информации.

В зависимости от конкретных обстоятельств имеющиеся сведения могут быть более полными или менее полными, и это весьма существенно определяет тот арсенал средств, которыми статистик может воспользоваться для решения данной задачи, и, следовательно, определяет качество решений, которое он может гарантировать потребителю.

Начнем с крайнего случая. Если нам достоверно известно реализовавшееся значение θ, задача становится тривиальной: необходимость в эксперименте отпадает; нам остается лишь сообщить потребителю то, что известно нам, но по каким-то причинам неизвестно ему. Аналогичная ситуация возникает, если нам достоверно известны n и $\mu(\theta, n)$; при этом, однако, необходимо произвести измерение X и осуществить операцию, обратную μ^*.

Менее тривиальным является случай, когда мы имеем полную информацию об операторе μ и некоторые специфические сведения о природе процесса n, но не о его реализации. Поясним это простым примером. Предположим, ставится задача обнаружения сигнала в шуме. Пусть шум аддитивен (т. е. μ есть просто оператор сложения). Пусть, далее, известно, что в некотором интервале частот спектр шума равен нулю, а спектр сигнала отличен от нуля. Ясно, что, произведя изме-

---

* При решении задачи обращения могут возникнуть трудности, но эти трудности имеют нестатистическую природу.

рения X в этом интервале частот, мы можем с любой степенью достоверности заключить, присутствует ли обнаруживаемый сигнал в шуме. Тривиальность этого примера является лишь кажущейся. Нам ведь может быть неизвестно, в каком представлении и какие компоненты шума равны нулю, а сигнала — отличны от нуля. Данное требование может не выполняться для представления Фурье, но окажется выполнимым при разложении принимаемого сигнала, скажем, в ряд по функциям Бесселя или полиномам Лягерра, или еще по какой-то системе функций. Выразимся точнее: если математическая модель, в терминах которой формулируется статистическая задача, допускает нулевую вероятность ошибки (при проверке гипотез) или нулевую дисперсию оценки (при оценивании параметра), то задача называется сингулярной (вырожденной). Интересно, что в случае, когда для сигнала и шума известны функции корреляции, Пьерри [1] удалось найти достаточное условие сингулярности задачи *, а Илюхин [1] дал обобщение этого условия.

Другой неожиданный пример сингулярности привел Ковер [1], который показал, что задача о проверке гипотезы о рациональности вероятности P некоторого события против альтернативной гипотезы об иррациональности P является вырожденной.

Следующим и более типичным уровнем априорной информации является полное вероятностное описание пространств Θ и B с помощью априорного распределения вероятностей $P(\theta)$ и семейства распределений (также известных априори) $F(x_1, x_2, \ldots, x_N/\theta)$, для всех $\theta \in \Theta$. В этом случае задачи решаются с помощью байесовых процедур. Наиболее уязвимым местом байесовых задач является необходимость задания априорного распределения $p(\theta)$. Все обстоит хорошо, если априорной информации о Θ действительно достаточно для задания $p(\theta)$. Однако иногда пытаются привести задачу к байесовской и тогда, когда таких сведений нет. Делается это с помощью постулата Лапласа — Байеса, согласно которому при отсутствии каких бы то ни было сведений о свойствах пространства Θ следует брать в качестве $p(\theta)$ равномерное в Θ распределение. При более детальном рассмотрении (см., например, Д. Худсон [1], § 17) оказывается, что такой метод может привести к противоречиям.

---

* Это условие выражается теоремой Пьерри:

Пусть N_i и S_i, $i=1, 2, \ldots$, являются коэффициентами разложения Карунена-Лоэва для шума и сигнала соответственно. Введем обозначения $\lambda_i = E(N_i^2)$ и $B_M = \sum_{i=1}^{M} S_i^2/\lambda_i$. Тогда, если $B_M \to \infty$ при $M \to \infty$, задача обнаружения сингулярна. Теорема допускает обобщение и на некоторые случаи оценки параметров.

Необходимости в искусственном задании $p(\theta)$ на самом деле нет, поскольку в математической статистике хорошо развита теория процедур, опирающихся только на информацию о семействе функций $F(x_1, x_2, \ldots, x_N/\theta)$. Для этого уровня априорной информации классической статистикой разработаны специальные методы синтеза, не требующие знания $p(\theta)$. В частности, найдено, что оптимальными (в условиях данного уровня) являются процедуры, использующие для оценок функцию правдоподобия, а для тестов—отношение правдоподобия. В рамках этого уровня возможны, конечно, и другие методы вынесения статистических решений; обычно их использование вызвано соображениями большей простоты реализации, но по качеству получаемых решений они всегда проигрывают методам, основанным на функции правдоподобия.

Проблемы, возникающие на последних двух уровнях априорной информации (т. е. при задании распределений на Θ и B или только на B), являются предметом классической математической статистики. Ее результаты изложены в многочисленных и широко известных монографиях и учебниках. Нас же будет интересовать другая ветвь математической статистики, связанная со следующим уровнем априорной информации.

Дело в том, что требование априорного знания функций распределения $F(x_1, \ldots, x^N/\theta)$ оказывается далеко не всегда выполнимым. Практикам все чаще приходится сталкиваться со статистическими проблемами, в которых действительно трудно сказать что-либо о виде распределений заранее. Прежде всего это относится к количественным исследованиям в области биологии, экономики, социологии; встречаются такие ситуации и в технике. До тех пор, пока приходилось иметь дело со случайными явлениями типа тепловых шумов, классическая статистика обеспечивала успешное решение возникающих задач. Но постепенно становилось ясным, что в более сложных ситуациях случайные факторы столь многочисленны, столь сложно связаны и нестабильны, что реальные распределения нельзя считать известными. Типичный пример дает история изучения линий радиосвязи на рассеянии, в ходе которой был накоплен огромный материал по поведению принимаемых сигналов. Попытки описать статистику сигналов с помощью известных распределений (например, релеевского, логнормального и т д) привели лишь к тому, что можно говорить только о процентах времени, в течение которого сигнал подчиняется этим распределениям, и притом — лишь с некоторой точностью. А ведь разработчику аппаратуры требуется реализовать в приемнике некоторую статистическую процедуру, которая обеспечивала бы постоянную работу при заданной надежности связи. Другой

похожий пример. В машиностроении было принято считать, что возможные распределения погрешностей изготовления деталей на металлорежущих станках могут быть сведены к одному из трех типов: равномерному, треугольному или нормальному (А. Н. Малов [1], Б С Балакшин [1]). Подробные исследования (М. А Левин и др. [1]) показали, что лишь в 20% случаев действительные распределения могут быть отнесены к одному из этих трех типов

Итак, существует необходимость решения статистических задач при условии, что распределения, участвующие в них, неизвестны.

Наиболее успешной попыткой удовлетворить эту потребность, оставаясь в рамках параметрических моделей, является метод минимакса, приводящий в конце концов к нахождению распределения, в некотором смысле «наилучшего среди плохих» или «наихудшего среди хороших». Несмотря на явное расширение класса ситуаций, для которых остаются справедливыми суждения, полученные методом минимакса,— недостатки параметрического подхода проявляются и здесь: для рассмотренного параметрического класса распределений даются лишь некоторые граничные суждения, а главное — нет гарантий, что истинное распределение будет всегда в рассмотренном классе, и ничего не известно о том, что будет, если распределение выйдет из этого класса

Непараметрическая статистика с самого начала, в самой исходной модели предполагает, что функциональный вид распределений, участвующих в задаче, неизвестен Это, конечно, не значит, что нам вообще ничего не известно Всякая задача, а статистическая — в наиболее явном виде, является, по сути дела, задачей выбора одного элемента из некоторого множества различных элементов. В статистике в качестве таких элементов выступают статистические гипотезы, т. е. те или иные предположения о распределениях. На предыдущих уровнях априорной информации (соответствующих классической статистике) конкурирующие гипотезы задаются в виде конкретных функций распределения (или параметрических семейств таких функций); тем самым полностью заданы и различия между гипотезами На рассматриваемом же уровне — которому отвечает непараметрическая статистика — априорная информация сводится к заданию только различий* между конкурирующими гипотезами, сами же распределения, охватываемые той или иной гипотезой, не кон-

---

* Понятие «различия» между классами распределений будет конкретизировано в § 17

кретизируются Подытожим сказанное выше следующей таблицей

Таблица 1.2 1

Уровень априорной информации	Достоверные знания о θ или p и n	Задание распределений на Θ и B	Задание распределении только на B	Задание параметрических классов распределений	Задание только различий между распределениями
Характер статистических задач	Задачи тривиальны или сингулярны	Байесовы задачи	Задачи на правдоподобие или субоптимальные задачи	Минимаксные задачи	Непараметрические задачи

В заключение сделаем важные замечания. Ясно, что чем больше априорной информации использует статистическая процедура, тем выше качество выдаваемых ею решений. Это означает, что байесовы процедуры дают наилучшие решения; затем идут процедуры, основанные на функциях правдоподобия; непараметрические процедуры должны давать наиболее «слабые» решения. Следует, однако, иметь в виду два момента.

Во-первых, такое упорядочивание процедур по качеству справедливо лишь в том случае, если априорная информация верна. При неверной априорной информации картина резко меняется: чем меньше априорной информации заложено в процедуру, тем слабее ухудшает решения ее ложность. Именно поэтому непараметрические процедуры во многих случаях оказываются лучше остальных.

Во-вторых, даже при верной априорной информации ее неполное использование не обязательно намного ухудшает качество решения На протяжении книги мы не раз встретимся с такими случаями, когда непараметрические процедуры проигрывают параметрическим в эффективности лишь несколько процентов; иногда они оказываются асимптотически оптимальными (т. е. проигрыш стремится к нулю с ростом объема выборки); а иногда вообще дают наилучшие достижимые решения (см., например, § 8.4). Как это ни удивительно, но это — факт. Если к тому же учесть, что некоторые непараметрические процедуры намного проще соответствующих параметрических, то легко понять, какое значение имеет развитие методов непараметрической статистики

§ 1.3. О МАТЕМАТИЧЕСКОЙ ТРАКТОВКЕ НЕИЗВЕСТНОСТИ РАСПРЕДЕЛЕНИЯ И О ТЕРМИНОЛОГИИ НЕПАРАМЕТРИЧЕСКОЙ СТАТИСТИКИ

Как уже отмечалось в предыдущем параграфе, непараметрическая статистика занимается задачами, в которых фиксируется только тип различия между конкурирующими распределениями, сами же распределения считаются полностью неизвестными. В 1942 г. Вольфовиц [1] сделал попытку дать математическую модель полностью неизвестного распределения, указав, что такое распределение не может быть описано с помощью функции с конечным числом параметров, в отличие от того, как это делается в классической статистике. В связи с этим он и ввел термин «непараметрическая гипотеза».

Так слово «непараметрический» оказалось единственным термином, связанным с существом этого раздела статистики В первой же монографии, обобщающей полученные к концу 50-х годов результаты, Фрэйзер [1] (стр. 126) применил этот термин в новом смысле, определив «непараметрическую статистику» как «тот раздел теории статистических выводов, для которого пространство параметров не может быть представлено просто как подпространство действительного пространства конечной размерности». Тут-то и начались трудности, не изжитые до сих пор Ударение оказалось поставленным не на неизвестности распределения, а на том, что число его параметров бесконечно. Показательно, что Фрэйзер сам заметил неудовлетворительность этого определения, указав на существование примеров, когда типично «параметрическая» задача согласно данному определению относится к числу непараметрических. С другой стороны, распределение может быть известным не в формульном, а в табличном или графическом виде, и если попытаться его записать аналитически, число параметров тоже может оказаться бесконечным. Но поскольку распределения известны, ничто не мешает применять обычные «параметрические» методы, например, вычислять отношения правдоподобия и строить оптимальные тесты

На другое неудобство, связанное с неудачной терминологией, указывает Нётер ([1], стр. 2). «В некотором смысле слово «непараметрический» может ввести в заблуждение. Так, «непараметрические» методы могут использоваться для нахождения доверительных интервалов для «параметров», таких, как медиана распределения. Трудность состоит в том, что термин «непараметрический», используемый в статистической литературе, недостаточно точен... Как и в большинстве

других публикаций по этим вопросам, в нашей книге термин «непараметрический» останется расплывчатым»

Однако, несмотря на все свои недостатки, этот термин прижился и прочно вошел в научный обиход; сейчас вряд ли имеет смысл придумывать другие названия для непараметрической статистики, непараметрического теста, непараметрической гипотезы и т. д. Попробуем дать определения этих терминов; во всяком случае, именно в указанном ниже смысле они будут употребляться в данной книге. (При этом придется употребить некоторые понятия, вводимые в книге лишь впоследствии).

Непараметрическая задача — статистическая задача, в которой указываются только различия между классами распределений*. По крайней мере, один из этих классов состоит из подчиняющихся некоторым довольно общим ограничениям, а в остальном неизвестных распределений; поэтому данный класс не может быть сведен к параметрическому семейству функций Такой класс распределений называется непараметрической гипотезой**.

Непараметрическая статистика — ветвь математической статистики, занимающаяся рассмотрением непараметрических задач и связанных с ними теоретических проблем.

Непараметрические методы синтеза и анализа статистических процедур — совокупность методов статистики, не предполагающих знания функционального вида распределений.

Непараметрические процедуры — алгоритмы решения непараметрических задач.

Непараметрический тест — термин, употребляемый в литературе в нескольких смыслах. Во-первых, это процедура проверки гипотез, по крайней мере одна из которых является непараметрической. При непараметричности нулевой гипотезы тест называется непараметрическим только в том случае, если его уровень значимости одинаков для всех распределений, охватываемых гипотезой При непараметричности альтернативной гипотезы оказывается, что разные классы непараметрических в указанном выше смысле тестов обладают различными типами устойчивости функции мощности при смене распределений в рамках альтернативы.

---

* Понятие «различия» см в § 1 7

** Даже в превосходной книге Уилкса [1] автору пришлось по инерции говорить о «простой непараметрической» гипотезе (а при переводе потребовалось дать этому специальное дополнительное пояснение — сравни тексты § 14 2(а) в оригинале и в русском издании). В нашем же понимании непараметрическая гипотеза не может быть простой, так как простой гипотезе соответствует единственное, а следовательно, известное распределение.

В связи с этим появляются такие понятия, как устойчивые (robust), усиленно непараметрические (strongly distribution-free) и другие тесты, о которых речь пойдет в соответствующих главах книги.

Непараметрическая оценка параметра. В непараметрическом случае оценка «параметров» возможна, если параметр есть известный функционал от неизвестного распределения. Оценку этого функционала, полученную без предположения о типе распределения, будем называть непараметрической. Отметим, что одному функционалу может соответствовать несколько непараметрических оценок с различными свойствами (см. гл. VIII).

Непараметрический факт — свойство выборки (или ее преобразований), которое не зависит от функционального вида распределения генеральной совокупности

Кроме того, в русской статистической терминологии слово «непараметрический» иногда употребляется как синоним английского термина «distribution-free», поскольку его дословный перевод, «не зависящий от вида распределения», слишком громоздок для частого употребления в тексте

§ 1.4. ТИПЫ НЕПАРАМЕТРИЧЕСКИХ ЗАДАЧ

Как уже неоднократно подчеркивалось, самым характерным для непараметрической статистики является то, что в ее задачах распределения вероятностей считаются полностью неизвестными, а сами задачи формулируются в терминах только различий между классами или внутри класса неизвестных распределений. В предыдущем параграфе обсуждалось, как именно следует понимать неизвестность распределения. Обсудим теперь характер конкретных задач, решаемых непараметрической статистикой.

Прежде всего необходимо упомянуть сугубо непараметрическую задачу оценивания неизвестных распределений. Ее не следует смешивать с проблемой аппроксимации неизвестного распределения известными функциями, которая рассматривается в обычной статистике и которая в конечном счете сводится к оценке параметров этих функций. В непараметрической постановке эта задача формулируется следующим образом: задается достаточно широкий непараметрический класс распределений (например, класс всех непрерывных функций распределения или класс всех распределений, обладающих плотностью и т. д); требуется предложить процедуру, результатом которой является оценка функции распределения или плотности; конечно, требуется также найти статистические свойства этой оценки, ха-

рактеризующие ее качество Легко видеть, что задача требует улавливать любые различия между распределениями внутри заданного класса, причем эти различия вообще не конкретизируются.

Если же эти различия конкретизировать, то мы приходим к другому кругу задач, аналогичных по своей форме классическим задачам оценки параметров. Аналогия здесь опять чисто внешняя, так как о параметрах распределения в обычном смысле говорить нельзя, поскольку класс распределения непараметричен. Фактически оценивается не параметр распределения, а параметр различия между распределениями внутри заданного непараметрического класса Любой параметр различия выражается некоторым функционалом от распределений, выражающим наше понимание этого различия (см. § 17)

Третья категория непараметрических задач — проверка непараметрических гипотез — наиболее близка и по форме и по существу к задачам проверки гипотез в классической статистике. (Может быть, именно поэтому данный раздел непараметрической статистики и развит сегодня наиболее сильно)

Любая задача проверки непараметрических гипотез выглядит следующим образом. Из двух конкурирующих гипотез альтернатива всегда непараметрична, а нулевая гипотеза может быть либо простой, либо непараметрической. Поскольку по крайней мере одна гипотеза есть класс неизвестных распределений, различие между гипотезами задается в некотором общем виде, не связанном с конкретным видом функции распределения Требуется предложить процедуру (тест), результатом которой явилось бы решение об истинности одной из гипотез на основании предъявленной выборки (или нескольких выборок — при многовыборочных задачах)

Перечислим теперь основные непараметрические задачи проверки гипотез и их варианты, чтобы продемонстрировать, как именно задаются гипотезы и различия между ними (При этом мы дадим постановки этих задач лишь в одновыборочном варианте, имея в виду, что обобщение на многовыборочный случай делается очевидным образом).

Задача согласия Пусть задано известное непрерывное распределение $F(x)$. Из неизвестного распределения $G(x)$, принадлежащего классу всех распределений, берется выборка $x_1, x_2, \ldots, x_N$.

Конкурирующие гипотезы.

Нулевая гипотеза	H_0:	$F=G$	простая гипотеза
Альтернатива а)	H_1^+:	$F<G$	односторонние гипотезы
б)	H_1^-:	$F>G$	
в)	H_1:	$F\neq G$	двусторонняя гипотеза

Как видим, нулевая гипотеза проста, альтернатива в любом из вариантов непараметрична, различие между ними выражается односторонним или простым неравенством между F и G

Задача сдвига (расположения) Иногда нам известно, что интересующий нас фактор приводит к сдвигу распределения в ту или иную сторону (причем не обязательно только к сдвигу, но к сдвигу — обязательно) Направление сдвига может быть известным или неизвестным В таких обстоятельствах возникает задача обнаружения сдвига, называемая иногда задачей расположения или локализации

В простейшей постановке задача расположения формулируется в том случае, когда известно, что альтернатива сводится только к сдвигу, т е. $F(x/\theta)=F_0(x-\theta)$

Нулевая гипотеза H_0: $\theta=0$.

Альтернатива а) $H_1': \theta>0$ } односторонние гипотезы
б) $H_1'': \theta<0$ }
в) $H_1''': \theta\neq 0$ двусторонняя гипотеза

В других случаях может быть неизвестным, проявляется ли влияние исследуемого фактора только в сдвиге, но известно, что сдвиг может иметь место Поскольку распределения неизвестны, среди них иногда могут встретиться и не имеющие моментов (по которым тоже можно было бы судить о сдвиге); поэтому естественной мерой сдвига являются квантили того или иного уровня p Возможны следующие варианты задачи сдвига:

Нулевая гипотеза $H_0: F^{-1}(p)=x_p$

Альтернатива а) $H_1: F^{-1}(p)>x_p$ } односторонние
б) $H_1': F^{-1}(p)<x_p$ } гипотезы
в) $H_1'': F^{-1}(p)\neq x_p$ двусторонняя гипотеза

В односторонних задачах иногда рассматривается несколько более общая нулевая гипотеза

$$H_0:\ F^{-1}(p)\leqslant x_0,$$
$$H_1:\ F^{-1}(p)>x_0.$$

В некоторых случаях известно, что распределение симметрично относительно медианы; тогда задача сдвига может быть сформулирована с учетом этой информации:

H_0 $F^{-1}(0{,}5)=x_0$ и $F(x)$ симметрична относительно x_0

а) $H_1': F^{-1}(0{,}5)>x_0$, F — симметрична,

б) $H_1'': F^{-1}(0{,}5)\neq x_0$, F — симметрична.

Знание симметричности распределения можно использовать при выборе структуры теста.

Задача расположения и симметрии. В отличие от задачи расположения, в данной задаче альтернатива расширяется так, чтобы охватить как все сдвинутые, так и все несимметричные распределения:

$$H_0\colon\ F^{-1}(0{,}5)=x_0 \quad \text{и } F(x) \text{ — симметрична}$$
$$H_1\colon\ F^{-1}(0{,}5)\neq x_0 \text{ или } F(x) \text{ — несимметрична.}$$

По существу, это задача проверки симметричности распределения $F(x)$ относительно точки x_0.

Задача масштаба В некоторых случаях заранее известно, что исследуемый фактор приводит к изменению масштаба распределения. Если изменяется только масштаб, имеем альтернативу вида $F(x/\theta)=F_0(\theta\cdot x)$ и задачу следующего типа:

H_0: $\theta=1$

а) H_1': $\theta>1$ } односторонние гипотезы

б) H_1'': $\theta<1$ }

в) H_1''': $\theta\neq 1$ двусторонняя гипотеза

Если кроме изменения масштаба могут происходить какие-либо другие изменения распределения, а нас интересует только сам масштаб, необходимо ввести меру масштаба. По тем же причинам, что и при сдвиге, разумно использовать некую квантильную меру или меру типа размаха выборки. Обозначим выбранную меру через ρ Тогда имеем следующую задачу:

H_0: $\rho=\rho_0$

а) H_1': $\rho>\rho_0$ } односторонние гипотезы

б) H_1'': $\rho<\rho_0$ }

в) H_1''': $\rho\neq\rho_0$ двусторонняя гипотеза

Задача независимости возникает в тех случаях, когда необходимо проверить, являются ли компоненты некоторого случайного вектора статистически связанными. При этом для гипотез используется самое общее определение статистической независимости — через факторизуемость распределения:

$$H_0: F(x^{(1)}, x^{(2)},\ldots, x^{(k)})=F^{(1)}(x^{(1)})\cdot F^{(2)}(x^{(2)})\cdots F^{(k)}(x^{(k)})$$
для всех $x^{(1)}, x^{(2)},\ldots, x^{(k)}$,
$$H_1: F(x^{(1)}, x^{(2)},\ldots, x^{(k)})\neq F^{(1)}(x^{(1)})\ F^{(2)}(x^{(2)})\cdots F^{(k)}(x^{(k)}).$$

хотя бы для некоторых значений $x^{(1)}, x^{(2)},\ldots, x^{(k)}$.
(Индексы поставлены вверху для того, чтобы охватить случай, когда $x^{(1)}$ и $x^{(k)}$ могут иметь различную физическую природу, т. е. отличить выборку $(x^{(1)}, \ldots, x^{(k)})$ от выборки

$(x_1, \ldots, x_k)$ отсчетов одной и той же величины X) Если вид распределений задается — имеем параметрическую задачу; если же распределения считаются неизвестными — задача непараметрична.

Задача случайности Многие из результатов математической статистики получены в предположении, что выборка является чисто случайной, т. е. состоит из одинаково распределенных и независимых величин. Если есть основания сомневаться, является ли данная выборка действительно чисто случайной или если изучаемый фактор приводит к нарушению случайности выборки, то имеет смысл решить задачу случайности:

$$H_0 : F(x_1, x_2, \ldots, x_N) = \prod_{i}^{N} F(x_i),$$

$$H_1 : F(x_1, x_2, \ldots, x_N) = \prod_{i}^{N} F_i(x_i) \neq \prod_{i}^{N} F(x_i).$$

Эта задача перекликается с задачей независимости, однако упор здесь делается на одинаковую распределенность величин X_i. В некоторых случаях оказывается возможным априори указать, чем именно будут отличаться распределения $F_i(x_i)$; тогда мы имеем более частные задачи Например, если альтернатива может быть сформулирована так, что измеряемые величины $\{X_i\}$ выразятся как $X_i = \xi C_i + Y_i$, где $\{Y_i\}$ — независимые, одинаково распределенные величины, $\{C_i\}$ — известный набор констант, а ξ — так называемая константа регрессии, то следующая задача называется задачей регрессии

H_0: $\xi = 0$

а) H_1': $\xi > 0$ односторонняя гипотеза

б) H_1'': $\xi \neq 0$ двусторонняя гипотеза.

По ряду причин еще более частная задача, когда C_i является монотонной функцией номера i, получила специальное название задачи тренда, или задачи тенденции.

Очевидно, что перечисленные выше задачи не исчерпывают всех возможностей различения распределений; здесь указаны лишь задачи, достаточно изученные (правда, в разной мере) к настоящему времени. Укажем также, что эти задачи допускают двувыборочную и k-выборочную формулировки, т. е. аналогичные задачи могут быть поставлены для сравнения двух или более выборок между собой по тому или иному различию между их распределениями

Подчеркнем еще раз в заключение, что данные задачи допускают и «параметрический» подход, если вид распре-

делений известен; но нас будут интересовать лишь непараметрические постановки, при которых распределения считаются неизвестными.

§ 1.5. ПРИНЦИП ИНВАРИАНТНОСТИ И РЕШЕНИЕ НЕПАРАМЕТРИЧЕСКИХ ЗАДАЧ

Было бы весьма желательно найти некоторый общий подход к решению каждой непараметрической задачи, т. е. найти такой алгоритм, действуя согласно которому, мы могли бы построить процедуру ее решения. Поскольку теперь распределения, участвующие в задаче, неизвестны, мы не можем записать функцию правдоподобия или функционал среднего риска и т. п., т. е. лишены возможности применять методы классической статистики. Необходимо найти метод, который давал бы решающие функции без предположения об известности распределений. Определенные надежды в этом отношении дает теория инвариантности, которая, по существу, является таким алгоритмом синтеза статистических процедур для задач, обладающих симметрией того или иного вида. Для более детального ознакомления с принципом инвариантности отошлем читателя к главе 6 книги Э. Лемана [1]; здесь же дадим лишь беглый обзор этапов синтеза процедур согласно этому принципу и обсуждение его пригодности для целей непараметрической статистики. Для определенности будем пока говорить лишь о задачах проверки гипотез.

Первый этап синтеза состоит в том, чтобы указать группу G преобразований g величины X, сохраняющих инвариантной задачу проверки гипотезы H_0 против альтернативы H_1. Дело в том, что многие статистические задачи обладают симметрией, математическим выражением которой и является инвариантность относительно некоторой группы преобразований. Например, если нулевая гипотеза состоит в том, что все $F_i(x)$, $i=1, 2, \ldots, N$, равны, а альтернатива в том, что не все они одинаковы, то* «естественно ограничиться рассмотрением только критериев [статистик], симметричных относительно значений $x_1, x_2, \ldots, x_N$, поскольку в противном случае принятие или отклонение той или иной гипотезы зависело бы (что совершенно не относится к делу) от нумерации этих переменных». В этом случае подходящей группой G преобразований g «является группа всех перестановок величин $x_1, \ldots, x_N$, поскольку функция от N переменных симметрична тогда и только тогда, когда она остается инвариантной при всевозможных перестановках этих переменных».

* Здесь и далее в этом параграфе в кавычках приводятся цитаты из главы 6 книги Э. Лемана [1].

Вторым этапом синтеза является отыскание максимального инварианта $T(x)$. Смысл этой операции состоит в следующем. Если к x применить последовательно все преобразования g из G, то получится совокупность эквивалентных точек, называемая траекторией G в выборочном пространстве. Функция $T(x)$ называется максимальным инвариантом, если она инвариантна и если она постоянна на каждой траектории, а на различных траекториях принимает различные значения. Например, пусть задана выборка $x_1, \ldots, x_N$, а G — множество $N!$ перестановок этих координат. «Тогда множество упорядоченных координат (порядковых статистик) $x_{(1)} \leqslant \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(N)}$ является максимальным инвариантом. Перестановка координат x_i не изменяет, очевидно, множества значений координат и поэтому не изменяет величин $x(\ )$ С другой стороны, две выборки с одними и теми же значениями порядковых статистик могут быть получены одна из другой перестановкой координат».

После того, как максимальный инвариант найден, задача редуцирована, т. е. сведена к совокупности новых случайных величин, сохранивших всю информацию о задаче в том смысле, что их статистические свойства при гипотезе и альтернативе различаются известным образом. Можно показать, что «класс всех инвариантных критериев совпадает со множеством критериев, зависящих только от максимальной инвариантной статистики $T(x)$». Поэтому следующий этап состоит в построении решающей функции δ на основе $T(x)$, чем и заканчивается синтез.

Если удастся пройти все эти три этапа без осложнений, то часто удается получить хороший (иногда даже равномерно наиболее мощный) критерий для решения данной задачи. Надо сказать, что принцип инвариантности иногда весьма небезуспешно претендует на решение непараметрических задач. Например, с его помощью очень красиво может быть построена теория ранговых тестов (см. Гаек и Шидак [1]). Подчеркнем, однако, те особенности принципа инвариантности, которые в известной мере ограничивают его практическое применение в непараметрической статистике.

Прежде всего, соображения инвариантности не являются универсальными, т. е. пригодными для решения любых задач. Для получения решающих процедур даже в рамках классической статистики его дополняют требованиями к другим свойствам критериев (несмещенности, почти инвариантности, некоторым свойствам функции мощности и т. д. — см. Леман [1]). В некоторых случаях принцип инвариантности вообще неприменим (см. Леман [1], стр. 310); иногда он применим, но его использование «не приносит успеха».

Во-вторых, использование принципа инвариантности даже в случае его успешного применения не обязательно приводит к получению непараметрических тестов, т. е. тестов, обладающих постоянством уровня значимости для всех распределений нулевой гипотезы.

И наконец, существенным недостатком этого принципа является его неполная алгоритмичность. В ряде случаев нахождение подходящей группы преобразований G не представляет труда, но нет четких рецептов построения такой группы в более сложных случаях Далее, нет также алгоритма нахождения максимального инварианта. Правда, иногда его можно «угадать»; в других случаях «оказывается удобным получать максимальные инварианты в несколько этапов, на каждом шаге находя их для подгрупп группы G». Однако Леман ([1], стр. 294) тут же приводит пример, показывающий, что успех на этом пути сильно зависит от выбора подгрупп, а главное — от того, в какой последовательности перебираются эти подгруппы. Никаких общих рекомендаций в этом отношении пока не имеется.

Все это вместе взятое приводит к необходимости развивать принципиально другие подходы к синтезу процедур решения непараметрических задач Один из таких подходов изложен в § 1 6, другой — в § 1 7

§ 1.6. ЭВРИСТИЧЕСКИЙ ПОДХОД К РЕШЕНИЮ НЕПАРАМЕТРИЧЕСКИХ ЗАДАЧ

Подход к решению непараметрических задач, излагаемый в данном параграфе, состоит в том, чтобы накопить арсенал средств, каждое из которых пригодно для решения конкретного типа (или нескольких типов) задач. Имея перед собой некоторую задачу, мы можем выбрать из этого арсенала конкретное «орудие» и приспособить его для решения этой задачи Подчеркнем, что при этом не имеется в виду пойти по пути простого накопления и рассортировки конкретных процедур, как это сделал Дж Уолш [1] в своем фундаментальном трехтомном «Справочнике по непараметрической статистике», хотя и такая работа чрезвычайно полезна Нас будут интересовать прежде всего принципиальные, теоретические вопросы получения результатов, пригодных для построения *классов* процедур, ориентированных на классы непараметрических задач; и только затем мы будем заниматься конкретными процедурами.

Во всех непараметрических задачах рассматриваются классы распределений, задаваемые лишь некоторыми общими для всех внешними свойствами, а в остальном распреде-

ления произвольны и неизвестны Такая общность в постановке задач наводит на вопрос: а не существует ли общности и в их решении? Оказывается, единство методики решения любых непараметрических задач существует и выражается в том, что каждая такая задача тем или иным способом сначала приводится к задаче, в которой фигурирует известное распределение Всегда сначала находится способ отображения непараметрической гипотезы (по крайней мере, одной, если их в задаче больше, чем одна) на простую, а затем уже так или иначе используется знание распределения, соответствующего этой простой гипотезе.

На первый взгляд может показаться странным, что множество неизвестных распределений оказывается возможным спроектировать на одно известное Следует, однако, иметь в виду, что неизвестность функционального вида распределения еще не означает, что нам абсолютно ничего не известно о свойствах выборки или некоторых функций от выборки. Задание даже самых общих ограничений на класс распределений приводит к появлению некоторых свойств, общих для выборок из любого распределения этого класса. Такие свойства выборки, не зависящие от вида распределения, называются непараметрическими фактами. Именно непараметрические факты и позволяют сводить непараметрические гипотезы к простым, и одной из основных задач непараметрической статистики является нахождение, изучение и систематизация таких фактов.

Можно выделить два основных типа непараметрических фактов (Ф. П. Тарасенко [2]). Первый — это непараметричность различных проявлений закона больших чисел: при достаточно широких ограничениях на свойства распределения выборки некоторые статистики обладают известными распределениями либо свойством сходимости их распределений к предельным и известным распределениям. Так, относительные частоты подчиняются биномиальному (и асимптотически нормальному) распределению; ряд линейных статистик имеет асимптотически нормальное распределение; другие статистики (например, статистики колмогоровского типа, крайние порядковые статистики и т. д.) обладают не нормальными, но известными асимптотическими распределениями, независимо от того, каково распределение выборки. (Необходимо помнить, что в ряде случаев при этом предполагается независимость выборочных значений).

Второи тип непараметрических фактов — это «внутренние» свойства самой исходной выборки, не зависящие от вида распределения. Метод вскрытия непараметрических свойств выборки сводится к тому, чтобы подвергнуть выбор-

ку некоторому преобразованию и исследовать статистические свойства преобразованной выборки. Во всех изученных до сих пор случаях оказывалось, что среди свойств преобразованной выборки всегда находились такие, которые сохранялись независимо от вида истинного распределения $G(x_1, ..., x_N)$ исходной выборки. Приведем (табл. 1.6.1) краткий перечень известных преобразований выборки и непараметрических свойств преобразованной выборки. (Заметим при этом что, по-видимому, нет ограничений на рассмотрение других преобразований и возможное установление новых непараметрических фактов).

Таблица 161

Тип преобразования	Непараметрические свойства
1 Перестановка элементов выборки	Равновероятность перестановок при симметричности $G(x_1, .., x_N)$.
2 Поэлементное приведение выборки в интервал [0,1] с помощью функции $F(x)$	Равномерность распределения приведенной выборки в интервале [0,1] при $F(x)=G(x)$
3 Упорядочивание элементов выборки по их величине	Сходимость R-й порядковой статистики по вероятности к квантилю уровня $R/(N+1)$: $x_{(R)} \to G^{-1}(R/(N+1))$.
4 Отображение на пространство выборочных блоков	Марковость и асимптотическая пуассоновость последовательности порядковых статистик.
	Статистическая эквивалентность выборочных блоков.
	Асимптотическая непараметричность функции мощности в классе распределений с одинаковой энтропией
5 Отображение на пространство ранговых векторов	Равновероятность ранговых векторов при симметрии $G(x_1, ., x_N)$.
	Статистическая связь рангов с соответствующими им выборочными значениями
6. Отображение на пространство ранговых интервалов	Чувствительность распределения ранговых интервалов к независимости и случайности элементов выборки*

Установив тот или иной непараметрический факт, мы можем указать класс задач, которые могут быть решены с его

* Пусть читатель не смущается тем, что в таблице могут встретиться незнакомые ему понятия или утверждения: все они будут разъяснены позже в соответствующих разделах книги.

помощью Например, свойства преобразований 2 и 4 могут использоваться для построения критериев согласия; 3 и 4 — для оценивания неизвестных распределений; 1, 5 и 6 — для проверки случайности или независимости выборки и т. д. Как именно использовать их, как строить конкретные процедуры — это вопрос, который мы рассмотрим впоследствии в главах третьей части книги.

Итак, методика построения процедур для решения непараметрических задач предстает теперь как последовательность выполнения следующих этапов.

1. Предлагается некоторое преобразование исходной выборки, результатом которого является новая выборка. Иногда это преобразование можно предложить, исходя из особенностей задачи, подобно тому, как это делается при использовании принципа инвариантности. Но можно делать это и эвристически, не привязываясь заранее ни к какой конкретной задаче, в расчете на то, что у преобразованной выборки удастся найти впоследствии какие-то непараметрические свойства. Пока неясно, существуют ли ограничения на класс возможных преобразований.

2. Изучаются статистические свойства преобразованной выборки в предположении, что распределение исходной выборки известно. Затем отыскиваются условия, при которых в окончательных формулах знание распределения несущественно, т. е. устанавливаются непараметрические факты. По-видимому, общим правилом является существование непараметрических фактов для любых разумных преобразований выборки; по крайней мере, до сих пор не найдено обратного примера.

3. Найденные непараметрические факты анализируются с целью нахождения адекватных им типов непараметрических задач. Например, статистическая эквивалентность выборочных блоков может использоваться для оценки неизвестного распределения, для построения критериев согласия, для процедур распознавания образов и, возможно, в каких-то других случаях. (Иногда даже если преобразование выборки предложено с учетом конкретной задачи, может оказаться, что непараметрические факты, связанные с этим преобразованием, имеют более широкое применение).

4. Предыдущие этапы выполняются, естественно, лишь теми, кто непосредственно занят развитием самой непараметрической статистики. Тот, перед кем стоит конкретная прикладная непараметрическая задача и кто нуждается в конкретной статистической процедуре для ее решения, имеет перед собой две возможности.

Первая состоит в выборе из числа известных процедур той, которая его по тем или иным причинам больше всего

устраивает Следует, однако, иметь в виду, что в ряде случаев мало что известно об относительных достоинствах разных процедур, предназначенных для решения одной и той же задачи (типичным является положение с многочисленными критериями согласия); поэтому здесь требуется известная осторожность и критичность В третьей части книги мы сделаем попытку собрать разбросанные по многочисленным статьям результаты по этим вопросам.

Вторая возможность заключается в применении разработанных к настоящему моменту методов синтеза непараметрических процедур к данной конкретной задаче. Надо сказать, что для ряда непараметрических процедур пока нет алгоритмических методов их синтеза, при их создании большую роль играют интуиция и эвристические методы. Однако во многих случаях найдены и алгоритмические методы. Это в особенности относится к синтезу тестов для проверки гипотез. В дальнейшем этот вопрос будет рассмотрен детально, а пока вкратце изложим общую методику синтеза непараметрических тестов.

Прежде всего, нулевая гипотеза должна быть представлена как простая. В некоторых задачах (например, в задачах согласия) нулевая гипотеза является простой в самой их постановке; если же нулевая гипотеза непараметрична, то она с помощью подходящего непараметрического факта проектируется на простую гипотезу. В итоге в любом случае мы имеем нулевую гипотезу, представленную единственным и известным распределением.

К сожалению, непараметрическая альтернатива остается непараметрической и после сведения нулевой гипотезы к простой. Поэтому обычно поступают следующим образом. Берется любое *известное* распределение, удовлетворяющее условиям альтернативы; известными из классической статистики методами синтезируется тест для проверки простой гипотезы против этой конкретной простой альтернативы; а затем этот тест используется для проверки гипотезы против любого из распределений альтернативы. Конечно, исторически многие тесты были предложены чисто эвристически: иногда по интуиции можно предложить конкретную статистику и предсказать, что она будет удовлетворять определенным требованиям к различению гипотезы от альтернативы Однако впоследствии почти всегда оказывалось, что каждый такой тест можно получить и указанным выше способом

Построенный таким образом тест обладает следующими свойствами. Во-первых, он будет непараметрическим в том смысле, что его уровень значимости не зависит от конкретного распределения даже в том случае, если нулевая гипотеза непараметрична. Причина этого лежит в представлении нуле-

вой гипотезы в виде простой. Во-вторых, хотя тест рассчитан на различение между нулевой гипотезой и данным «представителем» альтернативной гипотезы, он пригоден для проверки и против других распределений альтернативы Правда, при этом мощность теста будет зависеть от того, какое именно из распределений управляет выборкой; и, согласно фундаментальной лемме Неймана-Пирсона, мощность теста при этом может только уменьшаться Однако тест остается несмещенным; интересно, что в некоторых случаях существует целый подкласс распределений альтернативы, для которых мощность теста сохраняется постоянной (см. гл. 9).

Читатель, по-видимому, уже понял это, но мы все же подчеркнем еще раз, что в синтезе новых непараметрических процедур по данному методу пока еще большая роль отводится интуиции и изобретательности.

§ 1.7. МЕТОД ФУНКЦИОНАЛОВ В НЕПАРАМЕТРИЧЕСКОЙ СТАТИСТИКЕ

До сих пор мы рассматривали методы решения статистических задач, основанные на интерпретации статистик как функций выборочных значений. Этапы синтеза и анализа статистической процедуры при этом подходе выглядят следующим образом. Во-первых, основываясь на знаниях, опыте, интуиции, для каждой задачи можно построить или выбрать из известных такую функцию выборочных значений, которая будет чувствительной к различиям распределений, описанным в условиях задачи. Во-вторых, качество процедуры, основанной на предложенной статистике, можно определить, поскольку, зная функцию случайных величин и их распределения, в принципе всегда можно найти распределение самой функции, а по нему — все интересующие нас характеристики процедуры.

Практически всеми результатами современной статистики мы обязаны именно такому подходу. Однако по мере выдвижения практикой все более сложных задач и соответствующего развития теории стали проявляться трудности и недостатки, связанные с обоими этапами этого подхода. Первый этап — выбор подходящей статистики — носит эвристический характер, и в малоисследованных задачах статистику приходится буквально изобретать. В непараметрическом случае это приводит к типичной для изобретательства ситуации: всегда остается открытым вопрос, а нельзя ли придумать еще что-нибудь получше, все ли подходящие статистики мы перебрали. Второй этап — нахождение распределения статистики — в принципе преодолимый, поскольку вид статистики задан,

часто оказывается практически не приводящим к успеху из-за невозможности в сложных задачах довести до конца аналитические выкладки. Именно поэтому так велики трудности исследования мощностных свойств многих тестов, так мало известно о свойствах статистик при зависимых выборках и т. п. Статистическое моделирование на ЭВМ может пролить свет на свойства конкретной процедуры в конкретных условиях; это в ряде случаев устраивает практиков, но не может удовлетворить теоретиков из-за частности получаемых результатов.

Указанные трудности в значительной степени связаны с основной идеей подхода, согласно которой статистики рассматриваются как измеримые функции случайных величин. Р. Мизес [1] предложил принципиально иной подход, который позволяет преодолеть или обойти многие трудности исследования статистик. Его идея состоит в том, чтобы рассматривать статистику как функционал от эмпирической функции распределения. Эту мысль развивали В. И. Смирнов [6] и в особенности А. А. Филиппова [1], Чернов и Сэвидж [1], а также другие авторы. Ограниченное распространение, которое эта идея получила в статистике, объясняется прежде всего тем, что далеко не все статистики являются функционалами от эмпирической функции распределения (примерами, и не единственными, являются статистики на выборочных интервалах и крайних порядковых статистиках). Весьма существенное развитие идея Мизеса получила в исследованиях Хёфдинга [2], который рассматривал линейные интегральные функционалы от несмещенной оценки произведения маргинальных функций распределения. Класс получающихся при этом статистик назван им U-статистиками. Однако и этот класс еще не охватывает всех возможных статистик.

Тем не менее, уже на этом, пока еще частном, этапе развития данный подход явно демонстрирует свои преимущества как при синтезе статистических процедур, так и при их анализе. Действительно, вместо конструирования статистики теперь необходимо строить функционал, выражающий априорную информацию о задаче; статистика же получается автоматически, путем подстановки в функционал оценки функции распределения. Анализ же свойств статистики упрощается благодаря тому, что эти свойства могут теперь быть выражены через свойства оценки функции распределения.

Эти преимущества делают актуальным такое обобщение и развитие этого подхода, которое позволило бы охватить возможно более широкий класс статистик. Для этого нужно прежде всего обобщить интерпретацию статистики как оценки функционала.

Всякая задача проверки гипотез или оценивания параметра может рассматриваться как задача отбора (на основе результатов наблюдений) одного из подклассов некоторого класса распределений. Задачи проверки гипотез соответствуют дискретному (и обычно конечному) множеству подклассов; задачи оценки параметра — континуальному (в непараметрической задаче данному значению параметра всегда соответствует подкласс распределений, а в параметрической — либо подкласс, либо одно распределение, в зависимости от соотношения числа параметров семейства и числа оцениваемых параметров). Другими словами, в таких задачах необходимо на основе выборки провести различение подклассов Это достигается путем введения количественной характеристики различий между подклассами в виде функционала.

В непараметрической постановке задачи различие между распределениями обычно задается в настолько общем виде, что допускает неоднозначное толкование при формулировке количественной меры различия, что и приводит к возможности построения нескольких функционалов для одной и той же задачи. Проиллюстрируем это на простейшей задаче различения распределений по параметру расположения. Пусть требуется оценить параметр a, если задано только то, что он входит в неизвестное распределение $F_a(x)$ следующим образом: $F_a(x)=F(x-a)$. Как видим, указанное различие носит характер сдвига, но сам сдвиг еще не определен. Мы можем доопределить его как математическое ожидание ($a_1=\int x dF_a(x)$, как медиану ($a_2=\{x:F_a(x-a)=\frac{1}{2}\}$), как моду — при одномодовости распределений ($a_3=\{x:\max_x dF_a(x)/dx\}$), как полусумму симметричных квантилей ($a_4=\frac{1}{2}(x_p+x_{1-p})$, $\{x_q: F_a(x)=q\}$) и т. д. Таким образом, заданное различие (в данном случае сдвиг) может быть выражено несколькими функционалами, каждый из которых может быть оценен по выборке.

Итак, сформулированное в постановке задачи различие между распределениями конкретизируется с помощью задания некоторого функционала. При этом функционал не обязательно должен быть явно определен только на функциях распределения; в него могут входить плотности, производные плотности, обратные функции распределения, характеристические функции и другие представления распределений.

Конкретизировав заданное различие, то есть выбрав некоторый функционал, определенный на тех или иных представлениях распределений, участвующих в задаче, мы можем

перейти к синтезу статистики. Это достигается путем подстановки в функционал вместо входящих в него представлений распределений — оценок (в нашем случае — непараметрических) этих представлений. Так, если функционал имеет вид $J = \int \varphi(F, f, ..., f^{(r)}) dF$, то статистика есть $J_N = \int \varphi(F_N, f_N, ..., f_N^{(r)}) dF_N$, где φ — некоторая функция; F, f, ..., $f^{(r)}$ — функция распределения, плотность, ..., r-я производная плотности; F_N, f_N, ..., $f_N^{(r)}$ — их (непараметрические) оценки, соответственно, по выборке объема N.

Многообразие статистик, подходящих для решения данной задачи, объясняется не только тем, что заданное различие может быть выражено обычно не единственным функционалом, но и тем, что одно и то же представление распределения может быть оценено, как правило, несколькими способами. Например, плотность в непараметрическом случае может быть оценена гистограммой, оценкой Розенблатта-Парзена, полиграммой любого порядка, рядами по ортогональным функциям и т. д. Использование каждой из таких оценок для оценивания функционала будет давать новую статистику, со своими особенностями в свойствах*.

Итак, логически продолжая подход Р. Мизеса, можно дать следующее определение: статистика есть оценка некоторого функционала, в который вместо входящих в него представлений распределений, участвующих в задаче, подставлены статистические оценки этих представлений.

Предыдущие рассуждения позволяют надеяться на то, что это определение охватывает практически все известные статистики и может служить основой для синтеза новых статистик. Задача анализа теперь разбивается на два этапа: 1) изучение свойств оценок представлений распределений; 2) выражение свойств оценок функционала через свойства оценок фигурирующих в нем представлений. Декомпозиция сложной проблемы является мощным средством упрощения ее решения, и в данном случае тоже есть основания надеяться на проявление этого эффекта. Продемонстрируем это на простейшем примере.

Пусть нам необходимо оценивать линейный интегральный функционал $J = \int \varphi(\xi_1, ..., \xi_k)\ dF(\xi_1, ..., \xi_k)$. Его оценка выразится как $J_N = \int \varphi(\xi_1, ..., \xi_k)\ dF_N(\xi_1 ..., \xi_k)$. Выберем в качестве $F_N(\xi_1, ..., \xi_k)$ оценку, среднее которой равно

* В частности, поэтому не всегда сразу можно определить, какому функционалу соответствует данная статистика, построенная эвристически; иногда анализ такой «трудной» статистики может подсказать новую, ранее не исследованную оценку того или иного представления распределения.

$EF_N(\xi_1, \ldots, \xi_k)$ и ковариация которой равна $\operatorname{cov}[F_N(\xi_1, \ldots, \xi_k), F_N(\eta_1, \ldots, \eta_k)]$ Тогда (при определенных условиях регулярности, которые мы молчаливо будем предполагать, но пока не станем конкретизировать),

$$EJ_N = \int \varphi(\xi_1, \ldots, \xi_k)\, dEF_N(\xi_1, \ldots, \xi_k),$$

$$DJ_N = \iint \varphi(\xi_1, \ldots, \xi_k)\, \varphi(\eta_1, \ldots, \eta_k)\, d^{(2)} \operatorname{cov}[F_N(\vec{\xi}), F_N(\vec{\eta})].$$

Таким образом, если мы изучим свойства среднего и ковариации оценки F_N, мы можем получить суждения о среднем и дисперсии нашей оценки (при условии, что они существуют). Во многих случаях этой информации достаточно для практического использования предлагаемой оценки.

В данной книге мы ограничимся лишь формулировкой этой программы; лишь иногда (например, в гл. VIII) будут отражены некоторые ее аспекты. Ее реализация для одного класса интегральных (линейных и нелинейных) функционалов изложена в монографии под общей редакцией Ф. П. Тарасенко [8].

ГЛАВА II

ХАРАКТЕРИСТИКИ КАЧЕСТВА СТАТИСТИЧЕСКИХ ПРОЦЕДУР

§ 2.1. ВВЕДЕНИЕ

Для решения любой статистической задачи можно предложить несколько различных процедур. Выбор конкретной процедуры основывается на ряде соображений. Иногда большое значение придается простоте вычислений, требующихся для выполнения процедуры; в других случаях использование электронных вычислительных машин ослабляет это требование. При «ручных» вычислениях решающим может оказаться наличие достаточно подробных таблиц, связанных с той или иной процедурой. Однако при любых обстоятельствах нас прежде всего интересует качество решений, которые мы будем получать с помощью данной процедуры, т. е. их точность, надежность и т. д.

Для оценки качеств статистических процедур в математической статистике введены специальные понятия, разработаны различные методы сравнения процедур между собой. В данной главе мы дадим краткий обзор этих понятий и ме-

тодов. Наиболее детально рассмотрено сравнение тестов (§ 2.4 и далее), так как в последние годы в этом направлении получен ряд новых результатов.

§ 2.2. СВОЙСТВА ТОЧЕЧНЫХ ОЦЕНОК ПАРАМЕТРОВ (СТАТИСТИК)

Точечная оценка параметра — это некоторая статистика $\hat{\theta}=S(x_1, \ldots, x_N)$, которая используется вместо неизвестного истинного значения параметра θ. Ясно, что чем ближе $\hat{\theta}$ к θ, тем точечная оценка лучше. Но так как $\hat{\theta}$ является функцией выборочных значений, она оказывается случайной величиной, и, следовательно, как бы ни вводилось конкретное расстояние между $\hat{\theta}$ и θ, оно тоже будет случайным. Поэтому необходимо определить смысл «близости» между $\hat{\theta}$ и θ как меру соответствия между множествами их возможных значений. Такие меры могут быть различными, поскольку разные критерии близости могут уделять внимание различным качествам оценки.

Так как при неограниченном увеличении объема выборки мы обычно ожидаем неограниченного возрастания «близости» $\hat{\theta}_N$ к θ, то необходимо прежде всего определить точно, что мы будем понимать под сходимостью последовательности случайных величин $\hat{\theta}_N$ к θ при $N\to\infty$.

Говорят, что последовательность $\hat{\theta}_N$ ($N=1, 2, \ldots$) сходится к θ по вероятности, если для любого $\varepsilon>0$

$$\lim_{N\to\infty} Pr(|\hat{\theta}_N-\theta|\geqslant\varepsilon)=0. \qquad (2.2.1)$$

Для краткости утверждение о сходимости такого типа записывают в виде $p\lim\hat{\theta}_N=\theta$ или $\hat{\theta}_N\xrightarrow{p}\theta$.

Далее, говорят, что $\hat{\theta}_N$ сходится к θ в среднем, если

$$\lim_{N\to\infty} E(\hat{\theta}_N-\theta)^2=0. \qquad (2.2.2)$$

и $E(\hat{\theta}_N^2)$ и $E(\theta^2)$ ограничены (символ $E(X)$ означает математическое ожидание случайной величины X). Краткой формой записи этого определения является $l.i.m._{N\to\infty}\hat{\theta}_N=\theta$.

Из сходимости в среднем следует сходимость по вероятности. Действительно, согласно неравенству Чебышева

$$Pr\,(|\hat{\theta}_N-\theta|\geqslant\varepsilon)\leqslant\frac{E\,(\hat{\theta}_N-\theta)^2}{\varepsilon^2}, \qquad (2.2.3)$$

и из (2.2.2) следует (2.2.1).

Если оценка $\hat{\theta}_N$ сходится по вероятности к θ, то она называется состоятельной оценкой.

Если $E(\hat{\theta}_N/\theta)=\theta$, то оценка $\hat{\theta}_N$ называется несмещенной в среднем, чаще просто несмещенной. (Однако следует иметь в виду, что можно определить несмещенность и в другом смысле. Например, если условная функция распределения $F(\hat{\theta}_N/\theta)$ равна $^1/_2$ при $\hat{\theta}_N=\theta$, оценка называется несмещенной по медиане).

Смещенная оценка может быть состоятельной, если при $N\to\infty$ смещение $b^N(\theta)\to 0$, т. е. если оценка является асимптотически несмещенной (смещение определяется как $b_N(\theta)=E\hat{\theta}-\theta$.

Кроме величины смещения оценки ее качество определяется степенью разброса, рассеяния около среднего значения. Наиболее употребительной (но не универсальной) мерой рассеяния является дисперсия. Ее величину можно использовать как критерий качества оценки; так возникает понятие эффективной оценки. Если $\hat{\theta}_N$ — несмещенная оценка и имеет конечную дисперсию и для θ не существует другой оценки с меньшей дисперсией, то $\hat{\theta}_N$ называется эффективной оценкой параметра θ.

Естественно желание иметь количественную оценку наименьшей дисперсии $\hat{\theta}_N$. Такую возможность предоставляет известное неравенство Рао-Крамера (Г. Крамер [1]):

$$\sigma^2(\hat{\theta}_N)\geqslant\frac{(1+b'_N(\theta))^2}{I}, \qquad (2.2.4)$$

где I — количество информации по Фишеру,

$$I=E\left[\frac{\partial\ln p(x\,\theta)}{\partial\theta}\right)^2\Bigg].$$

Если эффективная оценка существует, то в (2.2.4) достигается знак равенства ($b_N(\theta)=0$), и дисперсия эффективной оценки обратно пропорциональна фишеровской информации.

Из теории известно (Крамер [1]), что равенство в (2.2.4) достигается при выполнении определенных условий:

а) регулярность функций, входящих в (2.2.4) (для существования производных и интегралов);

б) принадлежность распределения $p(x/\theta)$ к экспоненциальному семейству;

в) достаточность оценки $\hat{\theta}$ (свойство, которое мы обсудим ниже).

Таким образом, оказывается, что эффективность — свойство статистики по отношению к определенному классу распределений, что не для всех распределений существуют эффективные статистики — в смысле неравенства Рао-Крамера (т. е. обеспечивающие равенство в (2.2.4)). Может показаться, что каково бы ни было распределение $p(x/\theta)$, если несмещенные оценки существуют, найдется статистика с минимальной дисперсией, т. е. эффективная статистика; правда, у нас не всегда будет возможность оценить ее дисперсию явно. Однако можно привести пример, противоречащий этому «интуитивно обоснованному» утверждению (см. Э. Леман [1], гл. 1, задача 14). Кроме того, в ряде случаев для некоторых параметров может вообще не существовать несмещенных оценок (см. там же, гл. 1, задачи 4, 7, а также § 7.6).

Если нам известна дисперсия эффективной оценки $\sigma^2(\hat{\theta}_N)$, *то эффективность любой другой* (несмещенной) оценки $\tilde{\theta}^N$ того же параметра можно оценивать отношением

$$\mathrm{eff}(\theta_N, \tilde{\theta}_N) = \frac{\sigma^2(\hat{\theta}_N)}{\sigma^2(\tilde{\theta}_N)}. \qquad (2.2.5)$$

Ясно, что об эффективности в таком смысле можно говорить лишь при существовании эффективной оценки. Таким образом, понятие эффективности становится ограниченным, относительным *.

Иногда эффективная оценка данного параметра не существует, но может существовать асимптотически эффективная оценка $\hat{\theta}_N$, для которой равенство в (2.2.4) достигается при $N \to \infty$. В таком случае имеет смысл рассматривать *асимптотическую эффективность* точечной оценки $\hat{\theta}_\lambda$, определяемую как

* Определенный интерес представляет рассмотрение других мер эффективности (совокупные критерии, построенные на смещении и дисперсии; энтропийные или информационный критерии качества оценки и т. д. (Криц и Талако [1], Айрлэнд и Калбэк [1] и др.).

$$\mathrm{leff}(\hat{\theta}_N,\ \tilde{\theta}_N)=\lim_{N\to\infty}\frac{\sigma^2(\hat{\theta}_N|\theta)}{\sigma^2(\tilde{\theta}_N|\theta)}. \qquad (2.2.6)$$

Обратимся теперь к важному понятию достаточности оценки $\hat{\theta}_N$. Всегда желательно иметь статистику, наиболее сильно редуцирующую данные и в то же время сохраняющую всю полезную информацию о параметре θ. Приведем пример, показывающий, что такие статистики существуют. Пусть $(x_1, x_2, \ldots, x_N)$ — выборка независимых значений нормальной $N_x(\theta, 1)$ случайной величины X. Известно, что выборочное среднее $\bar{x}=\Sigma x_i/N$ является оценкой среднего θ. Можно ли улучшить эту оценку, подвергнув выборку дополнительной обработке? Легко показать, что распределение случайных величин $y_i=x_i-\bar{x}$, $i=1, 2, \ldots, N$ при заданном $\bar{x}$ не зависит от θ; это и означает, что из оставшихся после вычисления $\hat{\theta}=\bar{x}$ данных $y_1, y_2, \ldots, y_N$ нельзя больше извлечь никакой информации о параметре θ. В таких случаях говорят, что $\hat{\theta}_N$ содержит всю полезную информацию о θ.

По определению, статистика $\hat{\theta}$ называется достаточной оценкой для θ, если для любой другой оценки $\tilde{\theta}$ того же параметра условное распределение $p(\tilde{\theta}|\hat{\theta})$ не зависит от θ. В тех случаях, когда проверить достаточность оценки $\hat{\theta}$ по определению затруднительно, можно использовать так называемый критерий факторизации. Согласно этому критерию (необходимость и достаточность которого доказаны (см. Г. Крамер [1])), оценка $\hat{\theta}$ является достаточной тогда и только тогда, когда

$$p(x_1,\ldots, x_N|\theta)=p(\hat{\theta}|\theta)\cdot q(x_1,\ldots, x_N|\hat{\theta}),$$

где функция $q(x_1, x_2, \ldots, x_N|\hat{\theta})$ не зависит от θ.

Достаточность является более слабым требованием, чем эффективность: всякая эффективная оценка является достаточной, но не всякая достаточная — эффективной.

Кроме вышеперечисленных свойств точечных оценок иногда требуется охарактеризовать их другие качества. В случае необходимости тогда вводятся в рассмотрение дополнительные понятия и другие количественные соотношения.

Например, в непараметрической статистике внимание уделяется часто не столько тому, является ли статистика

$S(x_1, x_2, ..., x_N)$ оценкой какого-то параметра, сколько ее поведению при изменениях распределения выборки $G(x)$. Если окажется, что какие-то статистические свойства S_N при этом являются устойчивыми или известными, это используется при построении непараметрических процедур.

Так, в частности, появляется понятие непараметрической (distribution-free) статистики. Статистика $S_N(x_1, x_2, .., x_N)$ называется непараметрической в классе Ω распределений, если ее распределение вероятностей не зависит от того, какому распределению $G(x)$ из класса Ω подчиняется выборка.

Оказывается, что то, по отношению к какому классу Ω статистика S является непараметрической, существенно зависит от ее структуры. В связи с этим выдается дополнительная классификация статистик по их структуре.

Будем называть статистику $S(x_1, x_2, ..., x_N)$ статистикой структуры d, если она может быть представлена в виде некоторой симметричной функции W от величин $F(x_1)$, $F(x_2)$, ..., $F(x_N)$, где $F(x)$ — некоторая функция распределения (не обязательно та, которой подчиняется выборка x_1, x_2, ., x_N):

$$S(x_1,..., x_N)=W(F(x_1),..., F(x_N)). \qquad (2.2.7)$$

Статистики структуры d обладают тем полезным свойством, что если истинное распределение выборки $G(x)$ и функция $F(x)$, входящая в явное выражение (2.2.7) для статистики, совпадают, то распределение статистики не зависит от вида $F(x)$ $(=G(x))$ и может быть вычислено, если задан вид функции W*. Это просто объясняется тем, что преобразование $y=G(x)$ всегда дает величины, известным образом (равномерно) распределенные в [0, 1].

Введем, далее, в рассмотрение статистики структуры D. Для этого воспользуемся упорядоченной выборкой $x_{(1)}, x_{(2)}, ..., x_{(N)}$, в которой выборочные значения $x_1, x_2, ..., x_N$ расположены в порядке возрастания. Будем называть статистику $S(x_1, x_2, .., x_N)$ статистикой структуры D, если она представима в виде некоторой симметричной функции W величин $u_r=F(x_{(R)})-F(x_{(R-1)})$ $(R=1, 2,..., N, N+1; x_0=-\infty, x_{(N+1)}=\infty)$:

$$S(x_1,..., x_N)=W(u_1, u_2,..., u_{N+1}). \qquad (2.2.8)$$

Если $G(x)=F(x)$, то набор величин $u_1, u_2, ..., u_{N+1}$ является выборкой интервалов между равномерно распределенными

* Это дает некоторое основание делать встречающиеся в литературе утверждения типа «статистика структуры d является непараметрической (distribution-free) в классе Ω всех непрерывных распределений».

величинами, т. е опять-таки выборкой с известным распределением (см. § 4.2 гл. II), что также может быть использовано в ряде ситуаций.

§ 2.3. О КАЧЕСТВЕ ИНТЕРВАЛЬНЫХ ОЦЕНОК

Интервальное оценивание параметра состоит в том, что по выборке $(x_1, x_2, ..., x_N)$ вычисляются две статистики, $\underline{\theta}(\vec{x})$ и $\overline{\theta}(\vec{x})$, и знание интервала $(\underline{\theta}, \overline{\theta})$ используется в дальнейшем так, как если бы истинное значение параметра θ находилось в этом интервале.

Качество оценки теперь характеризуется двумя факторами: надежностью решения, которое выражается вероятностью того, что θ действительно находится в интервале $(\underline{\theta}, \overline{\theta})$; и «точностью» оценки, связанной с величиной самого интервала.

Можно, по-видимому, предложить различные интервальные оценки, но достаточно полно исследованы (и, следовательно, пригодны для использования в практике) оценки трех видов: доверительные, толерантные и фидуциальные* интервалы.

Если статистики $\underline{\theta}$ и $\overline{\theta}$ $(\underline{\theta}<\overline{\theta})$ могут быть выбраны так, что при заданном γ

$$\Pr[\underline{\theta}(\vec{x})<\theta<\overline{\theta}(\vec{x})|\theta]=\gamma,$$

то $\underline{\theta}$ и $\overline{\theta}$ называются нижней и верхней $100\gamma\%$-ными доверительными границами, $(\underline{\theta}, \overline{\theta})$ — доверительным интервалом, γ — доверительным коэффициентом.

Если статистики $\underset{\sim}{\theta}$ и $\tilde{\theta}$ $(\underset{\sim}{\theta}<\tilde{\theta})$ можно выбрать так, что распределение величины $F(\underset{\sim}{\theta})-F(\theta)$ (где $F(\hat{\theta}_N)$ — функция распределения $\hat{\theta}_N$) не зависит от вида функции распределения $F(x)$ и для $0<\beta<1$

$$Pr[F(\tilde{\theta}(\vec{x}))-F(\underset{\sim}{\theta}(\vec{x}))>\beta]=\gamma,$$

то $\underset{\sim}{\theta}$ и $\tilde{\theta}$ называются $100\beta\%$-ными независящими от распределения толерантными пределами уровня γ, а $(\underset{\sim}{\theta}, \tilde{\theta})$ — $100\beta\%$-ным толерантным интервалом уров-

* Относительно фидуциальных интервалов ограничимся лишь их упоминанием, интересующихся отошлем к книге С Уилкса [1]

ня γ. Вообще говоря, толерантный интервал не является оценкой некоторого численного параметра, от которого зависит функция распределения. Если нас интересует область, в которую с вероятностью не менее β попадет случайная величина X, то толерантный интервал является оценкой для такой области.

Надежность γ упомянутых выше интервальных оценок задается заранее и обеспечивается как выбором статистик для пределов (границ), так и достаточно большим объемом выборки. Зато величину интервала, по очевидным соображениям, в обоих случаях желательно минимизировать; а так как эта величина является случайной, то и подходы к ее минимизации могут быть различными (например, минимум средней величины интервала и пр.). Более детально эти вопросы будут рассмотрены в главе об оценках параметров.

§ 2.4. ТЕСТЫ, ИХ ХАРАКТЕРИСТИКИ И СВОЙСТВА

Важный класс статистических процедур образуют критерии проверки статистических гипотез, или, как их кратко называют, тесты. Статистической гипотезой называется всякое предположение о распределении вероятностей, характеризующем выборку. Ситуация обычно состоит в том, что с помощью некоторой процедуры (теста) необходимо сделать выбор между двумя конкурирующими гипотезами H_0 и H_1 (или, как часто говорят, между гипотезой H_0 и альтернативой H_1).

Всякий одновыборочный тест реализуется последовательностью следующих операций:

1. Получение совокупности выборочных значений $x_1, \ldots, x_N$.
2. Преобразование наблюденной выборки, $y_i = \psi_i(x_1, \ldots, x_N)$ *.
3. Вычисление значения заданной статистики $t = t(\vec{y})$.
4. Сравнение вычисленного значения t с заданным порогом (критическим уровнем) t_k и выбор соответствующего решения на основе результата сравнения по принципу: если $t < t_k$, принимается гипотеза H_0, если $t \geqslant t_k$, принимается альтернатива H_1.

Область T_0 значений t, удовлетворяющих неравенству $t < t_k$, называется областью принятия гипотезы, область T_1 значений $t \geqslant t_k$ называется критической областью (или областью отвержения гипотезы, или областью принятия альтернативы).

---

* Ясно, что если тест будет строиться на самой исходной выборке, то $\psi_i(x_1, \ldots, x_N) = x_i$.

С точки зрения потребителя решений качество процедуры определяется степенью соответствия между пространствами Θ и Γ (см. рис. 1.1.1). В случае тестов это соответствие можно выразить через вероятности ошибок. Если с гипотезами H_0 и H_1 связать параметры ситуации θ_0 и θ_1 соответственно, то вероятности ошибок выразятся как $\alpha = Pr(t \in T_1/\theta_0)$, $\beta = Pr(t \in T_0/\theta_1)$. Общепринято характеризовать тесты с помощью вероятности α, которая называется уровнем значимости теста, и вероятности $L(\theta_1) = 1 - \beta$, называемой мощностью теста.

Если множество Θ возможных значений параметра θ можно рассматривать как множество числовых значений, то тест можно охарактеризовать единой функцией $L(\theta)$, $\theta \in \Theta$, называемой оперативной характеристикой, или функцией мощности. Очевидно, $L(\theta_0) = \alpha$, $L(\theta_1) = 1 - \beta$.

Уровень значимости и мощность теста сами по себе еще недостаточно полно характеризуют тест, так как одних и тех же величин α и $L(\theta_1)$ можно достичь с помощью разных по качеству тестов просто за счет различия в объеме необходимых экспериментальных данных. Поэтому необходимым и естественным параметром качества разового теста является объем выборки N, соответствующий заданным уровню значимости и мощности. Для последовательного теста (у которого объем N выборки является случайным) полезной характеристикой может быть функция $E(N(\theta))$, выражающая зависимость среднего значения N от величины параметра θ.

Введя указанные характеристики тестов, мы получаем возможность сравнивать тесты между собой, говорить о свойствах и качествах тестов.

Например, если среди всех тестов, различающих θ_0 и θ_1 при одинаковом уровне значимости, найдется тест с максимальной мощностью, то он называется наиболее мощным тестом для данной альтернативы. Иногда оказывается, что если тест является наиболее мощным для некоторого значения θ_1, то он имеет наибольшую мощность и при всех остальных значениях θ; такой тест называется равномерно наиболее мощным критерием (сокращенно — РНМК); естественно, что найти РНМК для какой-то задачи проверки гипотез — это значит решить ее наилучшим образом.

Однако в ряде случаев РНМК либо вообще не существует, либо отыскать его в силу каких-то причин не представляется возможным. Тогда из всевозможных тестов выделяется тот или иной подкласс, в котором отыскивается тест, наилучший в каком-то смысле. Поскольку принцип выделения подкласса тестов и критерий оптимальности теста могут предлагаться

из различных соображений, на первый план могут выдвигаться различные свойства тестов. Перечислим наиболее употребительные свойства тестов, рассматриваемые как желательные в той или иной задаче.

Если тест имеет наибольшую мощность (при заданном уровне α) лишь для некоторого значения θ_1 и в его ближайшей окрестности ε, то такой тест называется *локально наиболее мощным* (ЛНМК).

Если для некоторого теста выполняется условие $\lim_{N\to\infty} L(\theta) = 1$, то тест называется состоятельным по отношению к альтернативе θ, или просто *состоятельным*. Это условие означает, что для состоятельного теста мощность может быть сделана сколь угодно близкой к единице за счет достаточного увеличения объема выборки.

Если выполняется условие $Pr(t \in T_1/\theta_0) \geqslant \alpha$ и $Pr(t \in T_1/\theta_1) > \alpha$, то критерий (тест) называется *несмещенным*. Несмещенность означает, что мощность теста превышает его уровень значимости, что, естественно, является желательным, но, к сожалению, не всегда возможным, в особенности в совокупности с требованием РНМК (Уилкс [1], стр. 405). Иногда смещенность теста вызывается просто тем, что областям T_0 и T_1 неверно приписаны индексы; в таких случаях достаточно изменить неравенства в операции 4) на противоположные.

Так как ЛНМК не является наиболее мощным для $\theta \in (\theta_1 \pm \varepsilon)$, полезно бывает оценить, насколько сильно падает мощность ЛНМК при удалении от точки θ_1. Так возникает понятие *критичности теста*, которое вводится следующим образом. Определим *огибающую функции мощности* тестов $\{t\}$ уровня α как $\varphi_\alpha(\theta) = \sup_{\{t\}} L_t(\theta)$. Тогда «нехватка» мощности теста t при альтернативе θ выразится как $\varphi_\alpha(\theta) - L_t(\theta)$; эта величина характеризует «недостаточность» теста t для альтернативы θ. Далее, величина $\Delta_t = \sup_{\{\theta\}} [\varphi_\alpha(\theta) - L_t(\theta)]$ выражает наибольшую потерю мощности данным тестом при всевозможных альтернативах. Теперь можно сравнивать Δ_t — характеристики для различных тестов из данного класса. Тест с минимальной величиной Δ_t называется *наименее критичным* (the most stringent) ЛНМК.

Перечисленные выше понятия, характеризующие свойства тестов, не исчерпывают всех используемых в статистике понятий. Так, говорят о монотонных, частично упорядоченных, подобных, инвариантных, перестановочных и др. тестах. Мы введем эти понятия позднее, по мере необходимости.

§ 2.5. СРАВНЕНИЕ ТЕСТОВ ПРИ КОНЕЧНОМ ОБЪЕМЕ ВЫБОРКИ

Итак, каждый тест характеризуется тремя основными величинами: уровнем значимости α, мощностью L и объемом выборки N. Очевидно, процедура 1 лучше процедуры 2, если

$$\alpha_1 \leqslant \alpha_2; \quad L_1 \geqslant L_2; \quad N_1 \leqslant N_2, \tag{2.5.1}$$

и по крайней мере одно из неравенств является строгим (см. рис. 2.5.1а). Однако гораздо чаще можно столкнуться с ситуацией, в которой два из неравенств (2.5.1) являются строгими, но в противоположных (2.5.1) направлениях (см. рис. 2.5.1 *б*, *в*, *г*. Графики а—г предполагают $N_1 = N_2$). Поэтому более полное суждение об относительных достоинствах двух тестов можно сделать лишь из сравнения их функций мощности в целом. Необходимо подчеркнуть, что явное вычисление функции мощности требует задания (или знания) а) параметров θ_0 гипотезы и θ_1 альтернативы, б) объема выборки N, в) уровня значимости α; и, что не всегда упоминается, но всегда подразумевается, г) распределений $p(x/\theta)$, $\theta \in \Theta$.

Таким образом, если нам удалось вычислить функции мощности $L_1(\theta, N, \alpha; p(x/\theta))$ и $L_2(\theta, N, \alpha; p(x/\theta))$ для двух тестов t_1 и t_2, то сравнение этих функций позволит полностью ответить на вопрос об относительных качествах t_1 и t_2: в тех областях переменных (θ, N), где $L_2 > L_1$, тест t_2 обладает лучшими

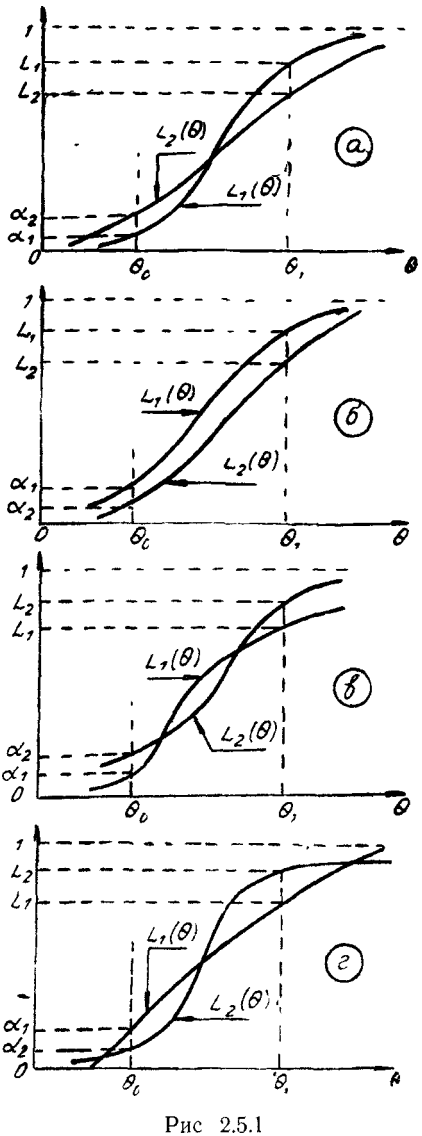

Рис 2.5.1

качествами, чем t_1 (при этом $p(x/\theta)$ и α считаются одинаковыми). Часто, однако, желательно знать не только то, что тест t_2 лучше t_1, но и насколько лучше.

Возможны различные подходы к количественному сравнению тестов. Один путь состоит в том, чтобы приравнять уровни значимости α_1 и α_2 и объемы выборок N_1 и N_2, а изменять, скажем, θ_2 до тех пор, пока мощности L_1 и L_2 не сравняются. Тогда отношение θ_1/θ_2 покажет, во сколько раз требуется изменить параметр альтернативы, чтобы t_1 и t_2 сравнялись по мощности; в некоторых прикладных задачах (типа обнаружения сигналов в шумах по параметру θ) такое сравнение имеет прямой инженерный смысл. Второй способ сравнения t_1 и t_2 состоит в вычислении так называемой питмановской относительной эффективности. Если дано два теста t_1 и t_2 одинакового уровня значимости α при одинаковой гипотезе θ_0, то относительная эффективность t_2 по сравнению с t_1 дается отношением N_1/N_2, где N_2 — объем выборки второго теста, необходимый для достижения той же мощности, которую имеет t_1 при одинаковой альтернативе и при объеме выборки N_1.

При сравнении тестов как с помощью первого показателя эффективности

$$e_1(t_1, t_2) = \left(\frac{\theta_1}{\theta_2}\right)_{N_1=N_2=N},$$

так и второго

$$e_2(t_1, t_2) = \left(\frac{N_1}{N_2}\right)_{\theta_1=\theta_2=\theta},$$

мы встречаемся с целым рядом трудностей. Во-первых, это чисто вычислительные трудности: очень часто аналитические выкладки при получении $L(\theta)$ в явном виде не доводятся до конца. Во-вторых, даже в тех случаях, когда e_1 и e_2 можно получить в аналитическом или численном виде, обе эти величины являются функциями α и θ (либо N) и функционалами от $p(x/\theta)$, что затрудняет обозримость результата сравнения, сужает его общность.

Трудности первого типа можно попытаться обойти, ограничившись рассмотрением класса тестов, основанных на асимптотически нормальных статистиках. В этом случае при не слишком малых, но конечных N функция мощности теста t с определенной степенью точности может быть записана как

$$L_t(\theta, N, \alpha;\ p_\theta(x)) = \Phi\left(\frac{t_{KN}(\alpha) - E_\theta(t_N)}{\sigma_\theta(t_N)}\right), \qquad (2.5.2)$$

где $\Phi(\lambda) = \int_\lambda^\infty \exp(-0{,}5\,x^2)\,dx/\sqrt{2\pi}$, $t_{KN}(\alpha)$ — порог срав-

нения для заданных N и α, E и σ — символы среднего и стандартного отклонения статистики t_N при заданном $p(x/\theta)$. Теперь показатели эффективности e_1 и e_2 тем же путем, что и раньше, могут быть найдены с помощью равенства

$$t\frac{{}_{1K_N}(\alpha)-E_{\theta_1}(t_{1N})}{\sigma_{\theta_1}(t_{1N})}=\frac{t_{2K_N}(\alpha)-E_{\theta_2}(t_{2N})}{\sigma_{\theta_2}(t_{2N})}. \qquad (2.5.3)$$

Но и при этом остаются трудности второго типа, т. е. остается недостаточная общность сравнения, слишком сильная связанность с конкретными условиями задачи.

§ 2.6. ТИПЫ АСИМПТОТИЧЕСКОГО ПОВЕДЕНИЯ ФУНКЦИИ МОЩНОСТИ

Желание упростить сравнение тестов, уменьшив число переменных функций $L(\theta, N, \alpha, p(x/\theta))$, приводит к мысли сравнивать тесты в некотором асимптотическом смысле. Идея состоит в том, чтобы сравнить поведение функций мощности двух тестов при неограниченном увеличении объема выборки N. Однако это оказывается нетривиальной задачей, так как асимптотическое поведение функции мощности зависит от того, как именно меняется порог t_{K_N} с ростом N и как ведет себя альтернатива. Варьируя эти факты, можно осуществить различные типы асимптотического поведения теста (табл. 2.6.1).

Таблица 2.6.1

Уровень значимости	Мощность	Альтернатива
$\alpha=\text{const}>0$	$L=\text{const}<1$	$\theta=\theta_N\to\theta_0$
$0<\alpha<1$, не фиксирован	$L\to 1$	$\theta=\text{const}\neq\theta_0$
$\alpha\to 0$	$L=\text{const}<1$	$\theta=\text{const}\neq\theta_0$
$\alpha=\text{const}>0$	$L\to 1$	$\theta=\text{const}\neq\theta_0$
$\alpha\to 0$	$L\to 1$	$\theta=\text{const}\neq\theta_0$

Соответственно будут различаться и меры относительной эффективности тестов. Мера, получаемая при условиях 1, получила название питмановской меры эффективности (см. Кендалл и Стьюарт [1], Нётер [2]). Бахадур ([1], [2] и др.) рассматривал вопросы сравнения тестов в условиях 2 и 3; любопытно, что случаи 3 и 4 неэквивалентны (Руист [1]); относительно поведения функции мощности в условиях 5 имеются пока лишь некоторые общие соображения (Бахадур [1], Сэвидж [1]). Задачей последующих параграфов данной главы является более детальное рассмотрение питмановской и

бахадуровской мер асимптотической относительной эффективности тестов. Мы постараемся подчеркнуть ограничения, которые приходится накладывать на тесты ради возможности их сравнить, и обсудим соотношения между различными мерами эффективности.

§ 27. АСИМПТОТИЧЕСКОЕ ПОВЕДЕНИЕ МОЩНОСТИ ТЕСТА (СЛУЧАЙ ПИТМАНА)

Согласно питмановскому подходу требуется сохранять постоянными величины $\alpha>0$ и $L(\theta)<1$ при $N\to\infty$. Очевидно, сделать это можно только в том случае, если обеспечить подходящее сближение альтернативы с гипотезой. Так мы приходим к необходимости рассматривать последовательность альтернатив θ_N, таких, что $\lim\limits_{N\to\infty}\theta_N=\theta_0$. Однако ясно, что не всякое стремление θ_N к θ_0 обеспечит постоянство α и L; нам предстоит теперь выяснить, при каких условиях это возможно.

Обратимся * к вычислению питмановской асимптотической мощности теста, т. е. величины

$$\lim_{\substack{N\to\infty\\ \theta_N\to\theta_0}} L(\theta, N, \alpha;\ p(x/\theta)).$$

Ограничимся сразу же классом асимптотически нормальных тестовых статистик t. Тогда при достаточно больших N критическая область, соответствующая уровню значимости α, задается соотношениями

$$t_N(\vec{x})\geqslant t_{K\alpha}(\alpha),\ \frac{t_{K\alpha}(\alpha)-E_{\theta_0}(t_N)}{\sigma_{\theta_0}(t_N)}=\lambda_\alpha,\ \varphi(\lambda_\alpha)=\alpha \qquad (2.7.1)$$

где $\varphi(\cdot)$ — интеграл вероятностей

Мощность теста при альтернативе θ_N запишется как

$$L(\theta_N, N, \alpha;\ p_{\theta_N}(x))=Pr\{t_N\geqslant t_{KN}|\theta_N\}=\varphi(\lambda_N), \qquad (2.7.2)$$

где

$$\lambda_N=[\sigma_{\theta_0}(t_N)\cdot\lambda_\alpha+E_{\theta_0}(t_N)-E_{\theta_N}(t_N)]/\sigma_{\theta_N}(t_N). \qquad (2.7.3)$$

Свяжем теперь последовательность альтернатив θ_N с N следующим образом

$$\theta_N=\theta+\frac{k}{N^\delta}, \qquad (2.7.4)$$

где k — произвольная положительная константа, а δ — кон-

* Будем следовать рассуждениям, проводимым Нётером в [2]

станта (тоже положительная), характеризующая скорость сходимости последовательности альтернатив к гипотезе. Так как теперь при $N\to\infty$ $\theta_N\to\theta_0$, то

$$E_{\theta_N}(t_N)=E_{\theta_0}(t_N)+\frac{1}{m!}\left(\frac{k}{N^{\delta}}\right)^m E^{(m)}_{\theta_N}(t_N)|_{\theta_N=\theta_0}+\ldots, \quad (2.7.5)$$

где m — порядок первой отличной от нуля производной от $E_{\theta_N}(t_N)$ по θ_N. Подставив (2.7.5) в (2.7.3) и ограничившись первым членом, получим:

$$\lambda_N=\frac{\sigma_{\theta_0}(t_N)\cdot\lambda_\alpha-\frac{1}{m!}\left(\frac{k}{N^{\delta}}\right)^m E^{(m)}_{\theta_N}(t_N)|_{\theta_N=\theta_0}}{\sigma_{\theta_N}(t_N)}. \quad (2.7.6)$$

Теперь, если дополнительно потребовать, чтобы

$$\lim_{N\to\infty}[\sigma_{\theta_0}(t_N)/\sigma_{\theta_N}(t_N)]=1, \quad (2.7.7)$$

$$\lim_{N\to\infty}[(N^{-m\delta}E^{(m)}_{\theta_N}(t_N)|_{\theta_N=\theta_0})/\sigma_{\theta_N}(t_N)]=C>0, \quad (2.7.8)$$

то

$$\lim_{N\to\infty}\lambda_N=\lambda_\alpha-\frac{k^m\cdot C}{m!}, \quad (2.7.9)$$

и

$$\lim_{\substack{N\to\infty\\ \theta_N\to\theta_0}}L(\theta, N, \alpha; p_\theta)=\varphi\left(\lambda_\alpha-\frac{k^m\cdot C}{m!}\right). \quad (2.7.10)$$

Здесь важно подчеркнуть, что, накладывая условие (2.7.8) на существование предела C, мы по необходимости связали скорость сходимости θ_N к θ_0, которая до сих пор выражалась произвольной константой, с определенными асимптотическими свойствами статистики t_N. В самом деле, пусть δ — константа, обеспечивающая существование предела C (2.7.8) и связанная, следовательно, с поведением статистики t_N. Пусть, далее, константа γ, характеризующая скорость сходимости θ_N к θ_0 (через соотношение $\theta_N=\theta_0+k/N^{\gamma}$, отличается от δ. Тогда

$$\lim_{N\to\infty}\lambda_N=\lambda_\alpha-\lim_{N\to\infty}\frac{k^m}{m!\,N^{(\gamma-\delta)m}}\cdot\frac{N^{-m\delta}E^{(m)}_{\theta_N}(t_N)}{\sigma_{\theta_N}(t_N)}. \quad (2.7.11)$$

Теперь видно, что если $\gamma\neq\delta$, то мощность теста стремится либо к α (при $\gamma>\delta$), либо к единице (при $\gamma<\delta$); в обоих случаях не соблюдается исходное требование постоянства L, т. е мы теряем возможность сравнения тестов.

§ 2.8. ПИТМАНОВСКАЯ АСИМПТОТИЧЕСКАЯ ОТНОСИТЕЛЬНАЯ ЭФФЕКТИВНОСТЬ

Определим питмановскую асимптотическую относительную эффективность A_{21} тестов, основанных на статистиках t_1 и t_2, как предел, к которому стремится отношение объемов выборки N_1/N_2 при $N\to\infty$, одинаковых α, равных мощностях и идентичных альтернативах $\theta_\Delta\to\theta_0$.

Очевидно, что если $\delta_1\neq\delta_2$, то результат будет либо нулем, либо бесконечностью, в зависимости от того, к какой из этих констант приравнять γ. Поэтому в дальнейшем будем предполагать, что сравнение всегда ведется в условиях $\delta_1=\delta_2=\delta$.

Согласно определению относительной эффективности по Питману, оба теста должны иметь одинаковые мощности при одинаковых альтернативах. Из (2.7.10) следует, что равенство мощностей предполагает

$$\frac{k_1^{m_1}\cdot C_1}{m_1!}=\frac{k_2^{m_2}\cdot C_2}{m_2!}, \tag{2.8.1}$$

а тождественность альтернатив (см. (2.7.4)) требует, чтобы

$$\frac{k_1}{N_1^{\delta}}=\frac{k_2}{N_2^{\delta}}. \tag{2.8.2}$$

Отсюда

$$\frac{N_1}{N_2}=\left(\frac{k_1}{k_2}\right)^{\frac{1}{\delta}}=\left(\frac{C_2}{C_1}\cdot\frac{m_1!}{m_2!}\cdot k_2^{m_2-m_1}\right)^{\frac{1}{m_1^{\delta}}}. \tag{2.8.3}$$

Как видим, правая часть (2.8.3) является положительной константой, которая и выражает относительную асимптотическую эффективность. Однако и тут нас ожидает подводный камень в виде произвольной константы k_2. Только в случае $m_1=m_2=m$ исчезает всякий произвол, и мы имеем

$$A_{21}=\lim_{N\to\infty}\frac{N_1}{N_2}=\left(\frac{C_1}{C_2}\right)^{\frac{1}{m\delta}}=\lim_{N\to\infty}\left\{\frac{E^{(m)}_{\theta_N}(t_2)\big|_{\theta_N=\theta_0}/\sigma_{\theta_0}(t_2)}{E^{(m)}_{\theta_N}(t_1)\big|_{\theta=\theta_0}/\sigma_{\theta_0}(t_1)}\right\}^{\frac{1}{m\delta}}=$$

$$=\lim_{N\to\infty}\frac{R_2^{\frac{1}{m\delta}}(\theta_0)}{R_1^{\frac{1}{m\delta}}(\theta_0)}, \tag{2.8.4}$$

где

$$R_i(\theta)=[E^{(m)}_{\theta}(t_i)/\sigma_{\theta}(t_i)]$$

называются константами эффективности i-го теста.

Как видим, при использовании для сравнения тестов асимптотической относительной эффективности по Питману мы исключаем зависимость результата от объема выборки N, от альтернативы θ (за счет соответствующих предельных переходов), и, кроме того, исчезает зависимость от уровня значимости α (за счет выпадения α из рассмотрения при переходе от (2.7.10) к (2.8.1)). Фактически остается только зависимость от распределения $p(x/\theta)$ через производную $E_\theta^{(m)}(t)$, но эта зависимость представляется вполне естественной.

Следует, однако, всегда иметь в виду, что для получения A_{21} в его окончательном виде (2.8.4) нам пришлось использовать целый ряд предположений, которые фактически несколько ограничивают значение и употребительность этого показателя.

Все эти предположения разбиваются на два класса: ограничения на свойства статистик (в результате наложения этих ограничений сравнению подлежат лишь тесты определенных классов) и ограничения на условия сравнения тестов (которые не позволяют экстраполировать результаты сравнения за пределы этих условий). Перечислим еще раз указанные ограничения и кратко их обсудим.

Ограничения на свойства статистик

1. Рассматривается класс асимптотически нормальных статистик. Хотя это ограничение не слишком сильно само по себе, оно может быть еще более ослаблено. По существу, нас интересуют лишь условия, при которых аргументы функций мощности будут равны для определенной последовательности альтернатив. Анализ (Кендалл и Стьюарт [1]) показывает, что требование асимптотической нормальности может быть заменено требованием, чтобы предельное распределение статистики зависело от двух параметров, один из которых являлся бы функцией θ. Так, для χ^2-распределения с ν степенями свободы и параметром нецентральности $\lambda(\theta)$ ($\lambda(\theta_0)=0$, т. е. при гипотезе распределение является центральным) $E(t)=\nu+\lambda(\theta)$, $\sigma(t)=\sqrt{2\nu}$, и формула (2.8.4) для асимптотической эффективности сохраняет свой вид.

2. От сравниваемых статистик требуется определенная регулярность, т. е. разложимость $E_\theta(t)$ в ряд по θ, существование участвующих в выкладках производных и интегралов, существование определенных пределов (см. (2.7.7) и (2.7.8)). Обычно выполнение или невыполнение условий ре-

гулярности проверяется непосредственно в ходе вычисления констант эффективности.

3. Сравнению могут подвергаться лишь статистики, для которых $\delta_1=\delta_2$. К тому, что было сказано выше по этому поводу, можно добавить, что, по существу, это есть требование одинаковой скорости убывания дисперсии статистики при возрастании N. Ясно, что если дисперсия одной из статистик стремится к нулю быстрее, то основанный на ней тест будет заведомо лучше. Таким образом, требование $\delta_1=\delta_2$ фактически разбивает все тесты на классы по различной сходимости дисперсии к нулю; чаще всего $\delta=\frac{1}{2}$; однако встречаются случаи, когда $\delta=\frac{1}{4}$ и $\delta=1$. В литературе пока нет (насколько известно автору) соображений о возможности или невозможности других значений δ.

4. Требование $m_1=m_2$ в свою очередь разбивает все тесты на классы, внутри которых только и имеет смысл сравнивать тесты между собой: ведь различие порядков производных, отличных от нуля, выражает существенное отличие в поведении статистик при альтернативах, близких к гипотезе. Для наглядности заметим, что, согласно этому требованию, нет смысла, например, сравнивать когерентный ($m=1$) и некогерентный ($m=2$) детекторы в радиолокационных системах, так как это процедуры различных классов, и их сравнение будет содержать определенный произвол.

Ограничения на условия сравнения тестов по асимптотической относительной эффективности

1. Так как сравнение тестов ведется при выполнении условия $\theta_N\to\theta_0$, то все заключения, основанные на коэффициенте A_{21}, носят локальный характер. Говоря инженерным языком, показатель A_{21} характеризует сравнительные качества двух процедур при обнаружении «слабых сигналов»; экстраполировать выводы о преимуществе одного из тестов на случай «сильных сигналов» без дополнительных исследований было бы по крайней мере неосторожно.

2. Организация стремления θ_N к θ_0 специального вида, при котором $\gamma=\delta$, может вызвать наибольшие сомнения в достаточной общности сравнения тестов по коэффициенту A_{21}. Задавая $\gamma=\delta$, мы тем самым как бы «подгоняем» условия сравнения к свойствам самих сравниваемых тестов. Однако, как мы видели в § 2.7, это единственный способ получить заданное асимптотическое значение функции мощности.

3. Конкретное значение коэффициента A_{21} можно вычислить, лишь задавшись конкретным видом распределения $p(x/\theta)$. Асимптотическая относительная эффективность в общем случае может менять свое значение от некоторого $A_0 \geqslant 0$ до ∞ только за счет изменений $p(x/\theta)$. Интересно, что в отдельных случаях удается найти $A_0 = \min\limits_{\{p(x/\theta)\}} A_{21}$: так, для $A_{21}(w, t)$, где w — статистика Вилкоксона, а t — известная t-статистика, найдено (Ходжес и Леман [1]), что $A_0(w, t) = 0{,}864$ (при $p(x/\theta)$ из класса непрерывных распределений); для $A_{21}(f, t)$, где f — статистика Фишера-Иейтса, $A_0 = 1$ (Чернов и Сэвидж [1]).

§ 2.9. БАХАДУРОВСКАЯ МЕРА ЭФФЕКТИВНОСТИ ТЕСТОВ

Обратимся теперь к сравнению эффективности тестов в случае, когда альтернатива θ фиксирована (Сэвидж [1], Бахадур [1], [2]). Для определенности будем рассматривать последовательность тестовых статистик $\{t_N\}$ с критической областью справа (т. е. альтернатива принимается, когда $t_\Lambda \geqslant t_{K_N}$). Будем рассматривать только те тесты, для которых при любом N существует единственная функция $L_N(t) = Pr(t_N \geqslant t/H_0)$. Это означает, что либо нулевая гипотеза проста, либо тест непараметричен, т. е. статистика t_N одинаково распределена при любом распределении, входящем в нулевую гипотезу. Очевидно, для непрерывных $L_N(t)$

$$Pr(L_N(t_\Lambda) \leqslant x \mid H_0) = x, \tag{2.9.1}$$

а если $L_\Lambda(t_\Lambda)$ дискретна, но имеет непрерывный предел, то

$$\lim_{\Lambda \to \infty} Pr(L_\Lambda(t_N) \leqslant x \mid H_0) = x. \tag{2.9.2}$$

Для любого состоятельного теста выполняется условие

$$Pr(\lim_{N \to \infty} L_\Lambda(t_N) = 0/H_1) = 1, \tag{2.9.3}$$

и сравнение тестов можно свести к сравнению их скоростей сходимости величин $L_N(t_\Lambda)$ к нулю в (2.9.3).

Учитывая, что $L(t)$ есть уровень значимости теста при заданном пороге t, назовем величину $L_N(t_N)$ **достигнутым уровнем**. Благодаря монотонности функции $L_N(x)$ решающее правило $t_\Lambda \geqslant t$ эквивалентно правилу $L_N(t_\Lambda) \geqslant L_N(t)$; поэтому все суждения о тесте, основанном на статистике t_N, можно получить, исследуя тест на $L_N(t_N)$. (2.9.2) и (2.9.3) говорят о том, что при нулевой гипотезе достигнутый уровень распределен равномерно в [0, 1]. Рассмотрим подробно

статистические свойства достигнутого уровня при $N\to\infty$ и фиксированной альтернативе θ.

Оказывается, что убывание достигнутого уровня при $N\to\infty$ для широкого класса тестов носит экспоненциальный характер: существует такая положительная константа $C(\theta)$, что

$$Pr\left(\lim_{N\to\infty}\frac{1}{N}\cdot\log L_N(t_N)=-\frac{C(\theta)}{2}\,\middle|\,\theta\right)=1. \qquad (2.9.4)$$

Величину $C(\theta)$, которая характеризует скорость убывания достигнутого уровня, для краткости будем называть коэффициентом наклона (exact slope) последовательности t_N.

С помощью коэффициента наклона можно охарактеризовать время (объем выборки), необходимое для того, чтобы достигнутый уровень опустился ниже некоторого заданного порога δ. Определим этот (случайный) объем выборки $n(\delta)$ как

$$n(\delta)=\begin{cases} m, \text{ где } m\text{—наименьшее целое число,} \\ \quad \text{такое, что } L_N(t_N)\leqslant\delta \text{ для всех } N\geqslant m; \\ \infty, \text{ если такое } m \text{ не существует.}\end{cases}$$

(Ясно, что при выполнении (2.9.4) $n(\delta)$ будет конечным с вероятностью единица). Тогда (Бахадур [1]) $n(\delta)$ и $C(\theta)$ связаны между собой предельным соотношением

$$Pr\left\{\lim_{\delta\to 0} n(\delta)/\left[\frac{2}{C(\theta)}\cdot\log\frac{1}{\delta}\right]=1\,\middle|\,\theta\right\}=1. \qquad (2.9.5)$$

Сопоставим теперь (2.9.3) и (2.9.5). Можно перефразировать (2.9.3) следующим образом: для любого заданного δ (>0) существует такое N_δ, что

$$Pr\{L_N(t_N)<\delta \text{ для всех } N\geqslant N_\delta/\theta\}>1-\delta. \qquad (2.9.6)$$

Поэтому при достаточно малых δ из (2.9.5) имеем:

$$N_\delta=\frac{2}{C(\theta)}\cdot\log\frac{1}{\delta}\cdot[1+0(\delta)]. \qquad (2.9.7)$$

Итак, мы получили N_δ, неслучайную меру того, как быстро достигнутый уровень для данного теста станет меньше заданной величины. Если мы имеем два теста, обладающих всеми качествами, необходимыми для получения формулы (2.9.7), то бахадуровская эффективность $\{t_{1N}\}$ по отношению к $\{t_{2N}\}$ определяется как

$$B_{21}(\theta)=\frac{N\delta_2}{N\delta_1}=\frac{C_1(\theta)}{C_2(\theta)}. \qquad (2.9.8)$$

Отослав интересующегося деталями доказательства соотношений (2.9.4) и (2.9.5) к упомянутым работам Бахадура и Сэвиджа, обсудим некоторые общие вопросы, связанные с бахадуровской эффективностью.

Во-первых, возникает вопрос, какие тесты допускают их сравнение с помощью меры $B_{21}(\theta)$. Ответ сводится к следующим требованиям на асимптотическое поведение тестовых статистик: должны существовать такая константа $b(\theta)$ и такая непрерывная функция $f(t)$, что

$$\text{а) } Pr\left\{\lim_{N\to\infty}\frac{t_N}{\sqrt{N}}=b(\theta)|\theta\right\}=1, \tag{2.9.9}$$

$$\text{б) } \lim_{N\to\infty}\left[\frac{1}{N}\log L_N(\sqrt{N}t)\right]=-f(t). \tag{2.9.10}$$

Этим требованиям удовлетворяют, например, асимптотически нормальные статистики

Во-вторых, важно указать метод вычисления коэффициента наклона $C(\theta)$. Докажем сначала, что из (2.9.9) и (2.9.10) следует

$$C(\theta)=2f(b(\theta)). \tag{2.9.11}$$

Выберем два произвольных положительных числа b_1 и b_2 таких, что $b_1<b(\theta)<b_2$. Из (2.9.9) следует, что для достаточно больших N $b_1\leqslant\frac{t_N}{\sqrt{N}}\leqslant b_2$ и, в силу монотонности $L_N(t)$, $L_N(\sqrt{N}b_1)\leqslant L_N(t_N)\leqslant L_N(\sqrt{N}b_2)$, или, что то же самое, $\frac{1}{N}\log L_N(\sqrt{N}b_1)\leqslant\frac{1}{N}\log L_N(t_N)\leqslant\frac{1}{N}\log L_N(\sqrt{N}b_2.)$ Привлекая (2.9.10), имеем: $-f(b_1)\leqslant\frac{1}{N}\log L_N(t_N)\leqslant -f(b_2)$. Так как b_1 и b_2 произвольны, то $\frac{1}{N}\log L_N(t_N)\to -f(b(\theta))$. Сопоставив этот результат с (2.9.4), получаем (2.9.11).

Итак, вычисление бахадуровской эффективности сводится к вычислению предельных функций $b(\theta)$ и $f(t)$. Чтобы продемонстрировать технику этих вычислений и трудности, которые при этом возникают, рассмотрим простой пример.

Пусть $t_N=N^{-1/2}\Sigma x_i$; и при нулевой гипотезе все X_i независимы и распределены нормально со средним 0 и дисперсией 1. Таким образом, $\lim_{N\to\infty}L_N(t)=\Phi(-t)$, где Φ — функция Лапласа. Пусть альтернатива состоит в том, что X_i распределены нормально со средним $\theta>0$ и дисперсией σ^2. Тогда

$$Pr\left\{\lim_{N\to\infty}\frac{t_N}{\sqrt{N}}=\theta/\theta,\ \sigma^2\right\}=Pr\left\{\lim_{N\to\infty}\frac{\sum x_i}{N}=\theta/\theta,\ \sigma^2\right\}=1,$$

и, следовательно, $b(\theta)=\theta$. Даже в этом простом случае точное вычисление $f(t)$ трудно. Сделаем дополнительное упрощающее предположение: примем, что $L_N(t)=\Phi(-t)$. Тогда

$$\lim_{N\to\infty}\frac{1}{N}\log L_N(\sqrt{N}\,t_N)=\lim_{N\to\infty}\frac{1}{N}\Phi(-\sqrt{N}\,t)=$$

$$=\lim_{N\to\infty}\log\left\{\frac{1}{\sqrt{2\pi N t^2}}\exp\left(-\frac{Nt^2}{2}\right)\right\}^{\frac{1}{N}}=-\frac{t^2}{2}.$$

Отсюда

$$f(t)=t^2/2 \text{ и } C^a(\theta)=2f(b(\theta))=\theta^2.$$

Отметим, что мы вычислили не точный коэффициент наклона $C(\theta)$, а приближенный, $C^a(\theta)$. Вообще говоря, $C^a(\theta)\neq C(\theta)$, но $\lim_{N\to 0}C^a(\theta)=C(0)$, и это, вместе с простотой вычислений, оправдывает рассмотрение приближенного коэффициента наклона.

§ 2.10. О СООТНОШЕНИЯХ МЕЖДУ РАЗЛИЧНЫМИ МЕРАМИ ЭФФЕКТИВНОСТИ

В ходе исследований качества различных статистических процедур многие авторы предлагали специфические меры их эффективности. Полезный обзор различных мер эффективности, обсуждение их связи, сравнение их между собой читатель может найти в книгах С. Рао ([1], стр. 414—420), Кендалла и Стьюарта ([1], гл. 25) и других источниках. В данном параграфе мы отметим лишь некоторые особенности бахадуровской и питмановской мер эффективности.

Покажем (Бахадур [2]), что при определенных условиях

$$\lim_{\theta\to\theta_0}B_{21}(\theta)=A_{21}. \tag{2.10.1}$$

Пусть $\{U_N\}$ — такая последовательность статистик, для которой при любом θ выполняются следующие условия:

а) существуют условные среднее $\mu_N(\theta)=E(U_N/\theta)$ и дисперсия $\sigma_N^2(\theta)=\mathrm{Var}(U_N/\theta)$;

б) величина $V_N=(U_N-\mu_N(\theta))/\sigma_N(\theta)$ имеет асимптотически нормальное распределение с нулевым средним и единичной дисперсией;

в) для величины $\Delta_N(\theta)=(\mu_N(\theta)-\mu_N(\theta_0))/\sigma_N(\theta_0)$ су-

ществует предел $\lim\limits_{N\to\infty}[\Delta_N(\theta)/\sqrt{N}]=b(\theta)$, где $b(\theta)\neq 0$ при $\theta\neq\theta_0$;

г)

$$\lim_{N\to\infty}\frac{1}{\sqrt{N}}\cdot\frac{\sigma_N(\theta)}{\sigma_N(\theta_0)}=0;$$

д) существуют такие $k>0$ (k — четно) и $\lambda>0$, что $b^2(\theta)=\lambda(\theta-\theta_0)^k\cdot(1+o(1))$ при $\theta\to\theta_0$.

Пользуясь соотношениями предыдущего параграфа, можно показать, что последовательность $t_N=|V_N|$ имеет коэффициент наклона, равный $b^2(\theta)$. Поэтому, если $\{t_{1N}\}$ и $\{t_{2N}\}$ соответствуют двум последовательностям $\{U_{1N}\}$ и $\{U_{2N}\}$, удовлетворяющим условиям а)—д), то $B_{21}=(b_1/b_2)^2$. Из условия д) имеем:

$$\psi_{21}(\theta_0)=\lim_{\theta\to\theta_0}B_{21}(\theta)=\begin{cases}0, & k_1>k_2\\ \lambda_1/\lambda_2, & k_1=k_2,\\ \infty, & k_1<k_2\end{cases}\qquad(2.10.2)$$

С другой стороны, из условия в) следует, что

$$\psi_{21}(\theta_0)=\lim_{\theta\to\theta_0}\lim_{N\to\infty}[\Delta_N^{(1)}(\theta)/\Delta_N^{(2)}(\theta)]^2.\qquad(2.10.3)$$

Правая часть (2.10.3) может быть сопоставлена с формулой для питмановской эффективности. Например, при $k_1=k_2=2$ и дифференцируемости функций $\Delta_N^{(i)}$ (обозначим их производные через $\Lambda_N^{(i)}$), изменив порядок перехода к пределам в (2.10.3), получим:

$$\psi_{21}(\theta_0)=\lim_{N\to\infty}[\Lambda_N^{(1)}(\theta_0)/\Lambda_N^{(2)}(\theta_0)]^2.\qquad(2.10.4)$$

Взяв теперь питмановское предположение о сходимости последовательности θ_N к θ_0 так, что

$$\lim_{N\to\infty}\frac{\Lambda_N^{(i)}(\theta_N)}{\Lambda_N^{(i)}(\theta_0)}=1,\ i=1,\ 2\qquad(2.10.5)$$

при совпадении уровней значимости и мощностей сравниваемых тестов, т. е.

$$\lim_{N\to\infty}\beta_N^{(i)}(\alpha/\theta_N)=\beta_0,\ i=1,\ 2,\qquad(2.10.6)$$

из (2.10.4) и (2.10.5) получим:

$$\psi_{21}=\lim_{\theta\to\theta_0}B_{21}(\theta)=\lim_{N\to\infty}\left[\frac{\Lambda_N^{(1)}(\theta_N)}{\Lambda_N^{(2)}(\theta_N)}\right]^2=A_{21},\qquad(2.10.7)$$

что и требовалось доказать.

Интересно отметить также прямую связь между питмановской эффективностью A_{21} и коэффициентом корреляции ρ между t_1 и t_2: $A_{21}=\rho^2$ (Ван Иден [1]) — в том случае, если t_1 — оптимальная статистика. Этот факт довольно просто доказывается. Действительно, пусть тест 1 основан на оптимальной статистике t_1 (т. е. его константа эффективности R_1 максимальна). Пусть другой тест, основанный на статистике t_2, имеет константу эффективности R_2. Рассмотрим третий тест, основанный на статистике

$$t(\lambda)=\lambda\frac{t_1}{\sigma_1(\theta_0)}+(1-\lambda)\frac{t_2}{\sigma_2(\theta_0)},$$

где λ — произвольная постоянная. Константа эффективности для t равна

$$R(\lambda)=\frac{\lambda R_1+(1-\lambda)R_2}{(\lambda^2+(1-\lambda)^2+2\lambda(1-\lambda)\rho)^{1/2}}.$$

Так как t_1 — наилучший тест, то при любом λ, $R(\lambda)\leqslant R_1$, откуда, с учетом предыдущей формулы,

$$\lambda^2[R_2^2-R_1^2-2R_1(R_2-\rho R_1)]-2\lambda[R_2^2-R_1^2-$$
$$-R_1(R_2-\rho R_1)]+R_2^2-R_1^2\leqslant 0$$

при любом λ. Следовательно,

$$[R_2^2-R_1^2-R_1(R_2-\rho R_1)]^2-$$
$$-(R_2^2-R_1^2)[R_2^2-R_1^2-2R_1(R_2-\rho R_1)]\leqslant 0.$$

Это тождественно условию $R_1^2(R_2-R_1\rho)\leqslant 0$, то есть $\rho=\frac{R_2}{R_1}$; откуда при $m\delta=1/2$ имеем $A_{21}=\rho^2$.

Часть II

НЕПАРАМЕТРИЧЕСКИЕ ФАКТЫ СТАТИСТИКИ

ГЛАВА III

СВОЙСТВА ПОРЯДКОВЫХ СТАТИСТИК

§ 3.1. ВВЕДЕНИЕ; ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ

Первое преобразование исходной выборки $(x_1, \ldots, x_N)$, которое мы подробно рассмотрим с целью поиска непараметрических фактов, состоит в упорядочивании элементов выборки по их величине. Пусть, для определенности, это упорядочивание производится по возрастающей величине. Тогда имеем последовательность $(x_{(1)}, x_{(2)}, \ldots, x_{(N)})$, в которой $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(N)}$. Будем называть эту последовательность упорядоченной статистикой (ее часто называют также вариационным рядом); величины $\{x_{(R)}\}$ получили название порядковых статистик; их номера $\{R\}$ называются рангами. (Скобки у номеров порядковых статистик введены для отличия их от порядковых номеров выборочных значений. Иногда может оказаться удобным или необходимым располагать как информацией о ранге R выборочного значения, так и его порядковом номере i в исходной выборке. В таких случаях будем пользоваться обозначением R_i, что означает «R есть ранг i-го наблюдения в выборке»; при этом подразумевается, что N известно и фиксировано). Если известны пределы a и b, ограничивающие интервал возможных значений переменной X, то можно построить дополненную упорядоченную статистику $(x_{(0)}, x_{(1)}, x_{(2)}, \ldots, x_{(N)}, x_{(N+1)})$, где $x_{(0)} = a$, $x_{(N+1)} = b$; ясно, что один (или оба) из пределов a и b может быть и бесконечным.

Обсудим теперь некоторые сложности, возникающие при упорядочивании.

В одномерном случае упорядочивание производится естественным образом, и трудность может возникнуть лишь если в выборке окажутся совпадающие по величине наблюдения*. Правда, это возможно только для дискретных случайных величин; в непрерывном же случае вероятность сов-

* Такие группы совпадающих выборочных значений будем называть связками.

падения двух или более выборочных значений равна нулю. Как упорядочивать выборочные значения при наличии связок — в значительной степени зависит от того, для чего производится упорядочивание; в гл. V этот вопрос встанет более конкретно, тогда мы и вернемся к нему. А пока лишь сочтем уместным заметить, что проблема связок достойна гораздо большего внимания со стороны теории, чем она пока получила*. Ведь все реальные измерения непрерывных величин производятся с *конечной точностью*, а следовательно, даже при высоких точностях измерений возможны связки, в особенности при больших объемах выборки.

В многомерном случае, т. е. когда элементами выборки служат случайные векторы, тоже может возникнуть необходимость упорядочивания. Делается это по-разному, опять-таки в зависимости от конкретных целей (см. гл. V), но если нам нужно приписать один порядковый номер каждой многомерной точке, т. е. иметь многомерный аналог упорядоченной статистики, то выход состоит в предварительном проектировании многомерной случайной величины на одномерную и последующем приписывании многомерным выборочным точкам номеров (рангов) их одномерных проекций. Выразимся более формально. Пусть m-мерная случайная величина $X_1 \cap X_2 \cap \ldots \cap X_m$ имеет непрерывную функцию распределения $F(x_1, x_2, \ldots, x_m)$. Пусть, далее, имеется m-мерная выборка объема N: $[(x_{11}, \ldots, x_{m1}), (x_{12}, \ldots, x_{m2}), \ldots, (x_{1N}, \ldots, x_{mN})]$. Введем упорядочивающую функцию $h(x_1, \ldots, x_m)$, с помощью которой любой совокупности значений $(x_1, \ldots, x_m)$ ставится в соответствие единственное число w: $h_1(x_{1K}, \ldots, x_{mK}) = w^K$. Наложив на функцию h требование, чтобы распределение величины w было непрерывным, мы получаем возможность упорядочить выборку $w_1, \ldots, w_K, \ldots, w_N$, и, следовательно, приписать соответствующие ранги многомерным точкам исходной выборки по следующей схеме:

$$\{(x_{1K}, \ldots, x_{mK})\} \to \{w_K\} \to \{w_{(R)}\} \to \{(x_{1(R)}, \ldots, x_{m(R)}\}.$$

Понятно, что такое упорядочивание неоднозначно, что при выборе различных функций h данный элемент одной и той же выборки может получить различные ранги; но такова уж цена за присвоение многомерным величинам одномерных ярлыков.

В заключение параграфа подчеркнем, что порядковые статистики имеют большое значение и в обычной параметрической статистике (см., например, Сархан и Гринберг [1], Дэвид [1]); нас же будут интересовать непараметрические свойства порядковых статистик.

* Исключением является, пожалуй, только книга Я. Гаека [1] по ранговым тестам.

§ 3.2. РАСПРЕДЕЛЕНИЕ УПОРЯДОЧЕННОЙ СТАТИСТИКИ И ПОРЯДКОВЫХ СТАТИСТИК

Рассмотрим статистические свойства упорядоченной статистики и порядковых статистик.

Совместное распределение всех порядковых статистик в случае независимых измерений получить легко. Для неупорядоченной выборки имеем (p и f — соответствующие плотности):

$$p(x_1, x_2,..., x_N)=p(x_1)\cdot p(x_2)... p(x_N), \ (\{x_i\in X\}, \quad i=1, 2,..., N). \tag{3.2.1}$$

После упорядочивания остается только информация о том, какие именно значения величины X были получены. Так как они могли быть получены в любом порядке и число возможных комбинаций (перестановок) равно $N!$, то

$$f(x_{(1)}, x_{(2)},..., x_{(N)})=N!\ p(x_{(1)})\cdot p(x_{(2)})... p(x_{(N)}), \quad x_{(1)}\leqslant x_{(2)}\leqslant ... \leqslant x_{(N)}. \tag{3.2.2}$$

Различные методы могут быть использованы для получения распределений отдельных порядков их статистик или любых их совокупностей. Приведем основные из этих методов.

1. Метод, использующий алгебру событий и основные правила теории вероятностей

Проиллюстрируем этот метод на примерах нахождения распределений наибольшей, наименьшей и любой (R-й) из порядковых статистик.

Событие $(X_{(N)}\leqslant x)$ эквивалентно пересечению событий $\{(x_i\leqslant x)\}$, $i=1, 2, ..., N$. Следовательно, по теореме умножения

$$G(x_{(N)})=Pr(X_{(N)}\leqslant x)=\prod_{i=1}^{N} Pr(X_i<x)=F^N(x), \tag{3.2.3}$$

где $G(x_{(N)})$ — функция распределения $X_{(N)}$, $F(x)$ — функция распределения X.

Отсюда, плотность вероятностей для $X_{(N)}$ равна

$$f(x_{(N)})=\frac{d}{dx_{(N)}} G(x_{(N)})=NF^{N-1}(x_{(N)})\cdot p(x_{(N)}). \tag{3.2.4}$$

Аналогично, поскольку $(x_{(1)}\geqslant x)\equiv\bigcap_i (x_i\geqslant x)$,

$$G(x_{(1)})=1-[1-F(x_{(1)})]^N, \tag{3.2.5}$$

и

$$f(x_{(1)})=N\,[1-F(x_{(1)})]^{N-1}\,p(x_{(1)}). \qquad (3.2.6)$$

Наконец, событие $(x_{(R)}\leqslant x)$ эквивалентно объединению событий «точно j значений из N измеренных меньше или равно x» для $j=R,\ R+1,\ R+2,\ \ldots,\ N$. Так как эти события взаимно несовместны, то

$$P_r(X_{(R)}\leqslant x)=\sum_{j=R}^{N}\binom{N}{j}F^j(x)\,[1-F(x)]^{N-j}. \qquad (3.2.7)$$

Следовательно, плотность вероятностей $x_{(R)}$ равна

$$f(x_{(R)})=\frac{d}{dx_{(R)}}\left\{\sum_{j=R}^{N}\binom{N}{j}F^j(x_{(R)})\,[1-F(x_{(R)})]^{N-j}\right\}=$$

$$=\frac{dF(x_{(R)})}{dx_{(R)}}\cdot N\cdot\sum_{j=R}^{N}\left\{\binom{N-1}{j-1}F^{j-1}(x_{(R)})\,[1-F(x_{(R)})]^{N-j}-\right.$$

$$\left.-\binom{N-1}{j}F^j(x_{(R)})\,[1-F(x_{(R)})]^{(N-1)-j}\right\},$$

что с учетом $\binom{N-1}{N}=0$ равно

$$\frac{N!}{(R-1)!(N-R)!}F^{R-1}(x_{(R)})[1-F(x_{(R)}]^{N-R}p(x_{(R)}). \qquad (3.2.8)$$

2. Метод, учитывающий, что вероятность попадания двух и более измерений в малый интервал имеет более высокий порядок малости, чем вероятность попадания только одного измерения в тот же интервал

Получим этим методом распределение (3.2.8). Вероятность того, что R-я порядковая статистика $X_{(R)}$ лежит в интервале $(x_{(R)}-\delta/2,\ x_{(R)}+\delta/2)$ при малых δ приближенно равна $f(x_{(R)})\cdot\delta$. С другой стороны, это событие можно считать эквивалентным тому, что

$R-1$ измерений меньше $x_{(R)}-\delta/2$,

одно измерение лежит между $x_{(R)}-\delta/2$ и $x_{(R)}+\delta/2$,

$N-R$ измеренный больше, чем $x_{(R)}+\delta/2$.

(Здесь мы и пренебрегаем вероятностью того, что в интервале δ может быть более одного измерения, так как эта вероятность порядка δ^2). Так как (опять приближенно!)

$$Pr(X<x_{(R)}-\delta/2)=F(x_{(R)})-\frac{1}{2}p(x_{(R)})\cdot\delta,$$

$$Pr(x_{(R)})-\delta/2<X<x_{(R)}+\delta/2)=p(x_{(R)})\cdot\delta,$$

$$Pr\,(X>x_{(R)}+\delta/2)=1-F(x_{(R)})-\frac{1}{2}p(x_{(R)})\cdot\delta,$$

то, используя полиномиальное распределение, получим:

$$f(x_{(R)})\cdot\delta=\frac{N!}{(R-1)!\,(N-R)!}F^{R-1}(x_{(R)})[1-F(x_{(R)})]^{N-R}p(x_{(R)})\cdot \cdot\delta+0(\delta^2).$$

Деля обе части равенства на δ и устремляя δ к нулю, получим формулу (3.2.8).

Этим же способом можно получить и другие распределения.

3. Метод интегрирования распределения упорядоченной статистики

Поскольку распределение (3.2.2) упорядоченной статистики является совместным распределением всех порядковых статистик, то распределение любой совокупности порядковых статистик можно получить интегрированием функции (3.2.2) по всем переменным, не входящим в интересующую нас совокупность.

Найдем, например, совместное распределение наибольшей и наименьшей порядковых статистик, $x_{(1)}$ и $x_{(N)}$.

$$f(x_{(1)}, x_{(N)})=\int_{x_{(1)}}^{x_{(N)}}\cdots\int_{x_{(1)}}^{x_{(4)}}\int_{x_{(1)}}^{x_{(3)}}f(x_{(1)},\ldots, x_{(N)})\,dx_{(2)}\ldots dx_{(N-1)}=$$

$$=N!\;p(x_{(1)})\,p(x_{(N)})\int_{x_{(1)}}^{x_{(N)}}p(x_{(N-1)})\int_{x_{(1)}}^{x_{(N-1)}}p(x_{(N-2)})\ldots$$

$$\ldots\int_{x_{(1)}}^{x_{(4)}}p(x_{(3)})\int_{x_{(1)}}^{x_{(3)}}p(x_{(2)})\,dx_{(2)}\ldots dx_{(N-1)}.$$

Последний интеграл вычисляется, и равен

$$\int_{x_{(1)}}^{x_{(3)}}p(x_{(2)})\,dx_{(2)}=F(x_{(3)})-F(x_{(1)}).$$

Далее,

$$\int_{x_{(1)}}^{x_{(4)}}[F(x_{(3)})-F(x_{(1)})]\,p(x_{(3)})\,dx_{(3)}=\frac{[F(x_{(4)})-F(x_{(1)})]^2}{2}.$$

Затем

$$\frac{1}{2}\int_{x_{(1)}}^{x_{(5)}}[F(x_{(4)})-F(x_{(1)})]^2\,p(x_{(4)})\,dx_{(4)}=\frac{[F(x_{(5)})-F(x_{(1)})]^3}{3!}$$

и так далее Окончательно,

$$f(x_{(1)}, x_{(N)}) = N(N-1)\,[F(x_{(N)}) - F(x_{(1)})]^{N-2}\,p(x_{(1)})\,p(x_{(N)}). \quad (3.2.9)$$

Более общо распределение совокупности s порядковых статистик $(x_{(k_1)}, x_{(k_1+k_2)}, \ldots, x_{(k_1+k_2+\ldots+k_s)})$ выборки из распределения $F(x)$ и плотностью вероятностей $p(x)$ выражается формулой

$$f(x_{(k_1)}, x_{(k_1+k_2)}, \ldots, x_{(k_1+k_2+\ldots+k_s)}) =$$

$$= \frac{N!}{(k_1-1)!\,(k_2-1)!\ldots(k_s-1)!\,(N-k_1-k_2-\ldots-k_s)!} \times$$

$$\times F^{k_1-1}(x_{(k_1)})\,[F(x_{(k_1+k_2)}) - F(x_{(k_1)})]^{k_2-1} \times \ldots$$

$$\ldots \times [1 - F(x_{(k_1+k_2+\ldots+k_s)})]^{N-k_1-k_2-\ldots-k_s} \times$$

$$\times p(x_{(k_1)}) \cdot p(x_{(k_1+k_2)}) \ldots p(x_{(k_1+k_2+\ldots+k_s)}). \quad (3.2.10)$$

Очевидно, $k_1, k_2, \ldots, k_s$ — целые числа и $k_i \geqslant 1$, $\sum_1^s k_i \leqslant N$ и

$$x_{(0)} \leqslant x_{(k_1)} \leqslant \ldots \leqslant x_{(k_1+k_2+\ldots+k_s)} \leqslant x_{(N+1)}.$$

Формула (3.2.10) впервые была получена Крэйгом [1].

4 Методы, использующие марковость порядковых статистик

Еще в 1933 году А. Н. Колмогоров [1] обратил внимание на то, что порядковые статистики образуют простой марковский процесс. Действительно, условное распределение вероятностей $(i+1)$-й порядковой статистики при заданной i-й порядковой статистике не зависит от значений статистик с номерами, меньшими i. В самом деле, вычислив с помощью (3.2.10) искомую условную вероятность, получим:

$$f(x_{(i+1)} | x_{(i)}, x_{(i-1)}, \ldots, x_{(1)}) = \frac{f(x_{(1)}, x_{(2)}, \ldots, x_{(i+1)})}{f(x_{(1)}, x_{(2)}, \ldots, x_{(i)})} =$$

$$= (N-i)\,[1 - F(x_{(i+1)})]^{N-i-1}\,[1 - F(x_{(i)})]^{-N+i}\,p(x_{(i+1)}) =$$

$$= f(x_{(i+1)} | x_{(i)}). \quad (3.2.11)$$

Таким образом, порядковые статистики $x_{(1)}, x_{(2)}, \ldots, x_{(N)}$ образуют неоднородный процесс Маркова с дискретным параметром, который характеризуется начальной мерой (3.2.5) и переходной вероятностью

$$Pr(X_{(i+1)} \leqslant x / x_{(i)} = y) = 1 - [1 - F(x)]^{N-i}\,[1 - F(y)]^{-N+i} \quad (3.2.12)$$

(последнее равенство легко получается путем соответствую-

щего интегрирования (3.2.11)). Поэтому многие выводы о свойствах порядковых статистик и функциях от них могут быть получены с помощью методов теории марковских процессов (см., например, Пайк [1], Реньи [1], М. Рао [1]).

§ 3.3. РАСПРЕДЕЛЕНИЕ ПОРЯДКОВЫХ СТАТИСТИК В ДИСКРЕТНОМ СЛУЧАЕ

Как мы уже отмечали, в дискретном случае в упорядоченной статистике возможно появление «связок» (то есть совпадение нескольких выборочных значений). Для простоты обратимся к случаю, рассмотренному Хатри [1], когда возможные значения X есть числа натурального ряда 0, 1, 2, ... с заданными вероятностями $p(0)$, $p(1)$, $p(2)$, ... Естественно, $p(l) \geqslant 0$ и $\sum_{l}^{\infty} p(l) = 1$. Если число возможных значений X конечно, будем считать, что остальные числа имеют нулевую вероятность.

Найдем распределение R-й порядковой статистики. (Подчеркнем, что вопрос о том, как присваиваются различные ранги внутри связки, остается открытым: это пока несущественно). Пусть $(R-1-k)$ значений меньше $x_{(R)}$, $(N-R-m)$ — больше него, остальные $(k+m+1)$, включая $x_{(R)}$, находятся в данной связке. Так как k и m могут иметь различные значения ($k=0, 1, 2, \ldots, R-1$; $m=0, 1, \ldots, N-R$), то возможны $R \cdot (N-R+1)$ различных взаимно несовместных комбинаций событий. Учитывая, что вероятности перечисленных выше трех событий равны соответственно

$$Pr(X<x_{(R)}) = F(x_{(R)}-1) = \sum_{l=0}^{x_{(R)}-1} p(l),$$

$$Pr(X>x_{(R)}) = 1 - \sum_{l=0}^{x_{(R)}} p(l) = 1 - F(x_{(R)}),$$

$$Pr(X=x_{(R)}) = p(x_{(R)}),$$

имеем окончательно

$$f(x_{(R)}) = \sum_{k=0}^{R-1} \sum_{m=0}^{N-R} \cdot$$

$$\cdot \frac{N!\,[F(x_{(R)}-1)]^{R-1-k} p^{k+m+1}(x_{(R)})[1-F(x_{(R)})]^{N-R-m}}{(R-1-k)!\,(k+m+1)!\,(N-R-m)!}. \quad (3.3.1)$$

Используя соотношение

$$\int_0^1 y^k (1-y)^m dy = \frac{k!\,m!}{(k+m+1)!}, \quad (3.3.2)$$

можно привести (3.3.1) к более обозримому виду. Действительно, перепишем (3.3.1) в форме

$$f(x_{(R)})=R\binom{N}{R}\sum_{k=0}^{N-R}\sum_{m=0}^{N-R}\binom{R-1}{k}\binom{N-R}{m}]F(x_{(R)}-1)]^{R-1-k}\times$$

$$\times[1-F(x_{(R)})^{N-R-m}p(x_{(R)})\int_0^1[yp(x_{(R)})]^k[(1-y)p(x_{(R)})]^m dy. \tag{3.3.3}$$

Изменив порядок суммирования и интегрирования и производя замену $yp(x_{(R)})+F(x_{(R)}-1)=w$, получим

$$f(x_{(R)}=R\binom{N}{R}\int_{F(x_{(R)}-1)}^{F(x_{(R)})}w^{R-1}(1-w)^{N-R}dw. \tag{3.3 4}$$

Выражение для функции распределения $G(x_{(R)})$ получится в виде:

$$G(x_{(R)})=\sum_{x_{(R)}-0}^{x_{(R)}}f(x_{(R)})=R\binom{N}{R}\int_0^{F(x_{(R)})}w^{R-1}(1\quad w)^{N-R}dw. \tag{3 3 5}$$

Интегралы, входящие в (3.3.4) и (3.3.5), выражаются через неполные бета-функции, таблицы или ряды для которых можно использовать в случае необходимости конкретных расчетов.

Аналогичным способом можно получить совместное распределение $x_{(R)}$ и $x_{(L)}$ $(L>R)$:

$$f(x_{(R)}, x_{(L)})=\sum_{k=0}^{R-1}\sum_{m=0}^{N-L}\sum_{u,t}'\frac{N![F(x_{(R)}-1)]^{R\ 1-k}p(x_{(R)})^{1+k+t}}{(R-1-k)!\,(1+k+t)!(L-R\ \ 1-u-t)!}\times$$

$$\times\frac{[F(x_{(L)}-1)-F(x_{(R)})]^{L-R-1-u-t}p(x_{(L)})^{1+m+u}[1-F(x_{(L)})]^{N-L-m}}{(1+m+u)!\ \ (N-L-m)!}, \tag{3.3.6}$$

где $\sum_{u,t}'$ обозначает суммирование по u, $t=0, 1, \ldots, L-R-1$, причем такое, что $(L-R-1)\geqslant(u+t)$. Тем же приемом, что и выше, (3.3.6) может быть приведено к виду

$$f(x_{(R)}, x_{(L)})=C\int\int w^{R-1}(v-w)^{L-R-1}(1-v)^{N-L}dv\,dw, \tag{3.3.7}$$

где $C=N!/(R-1)!(L-R-1)!(N-R)!$ и интегрирование ведется по области

$$v\geqslant w,\ F(x_{(R)})\geqslant w\geqslant F(x_{(R)}-1),\ F(x_{(L)})\geqslant v\geqslant F(x_{(L)}-1).$$

Далее, можно показать (Хатри [1]), что

$$E(x_{(R)}) = \sum_{i=0}^{\infty} [1 - Q_R(i)], \qquad (3.3.8)$$

и

$$E(x_{(R)}^2) = \sum_{i=0}^{\infty} (2i+1)[1 - Q_R(i)], \qquad (3.3.9)$$

где $Q(i) = I_{\Gamma(i)}$ $(R, N-R+1)$ — неполная бета-функция.

§ 3.4. АСИМПТОТИЧЕСКИЕ РАСПРЕДЕЛЕНИЯ КРАЙНИХ ПОРЯДКОВЫХ СТАТИСТИК

Рассмотрим статистические свойства порядковых статистик при неограниченном возрастании объема выборки. Для простоты начнем со случая равномерно распределенных случайных величин. В этом случае (3.2.8) запишется в виде

$$f(x_{(R)}) = \frac{N!}{(R-1)!\,(N-R)!} x_{(R)}^{R-1} (1 - x_{(R)})^{N-R}. \qquad (3.4.1)$$

Теперь, если $N \to \infty$, следует различать два случая: когда $x_{(R)}$ близко к крайним значениям выборки, и когда $x_{(R)}$ находится в среднем диапазоне выборочных значений, поскольку предельные распределения для этих случаев различны.

В данном параграфе мы будем рассматривать предельные распределения порядковых статистик, отстоящих на фиксированном месте R от начала упорядоченной статистики. Из (3.4.1) легко получить распределение величины $y_R = N x_{(R)}$:

$$f(y_R) = \frac{(N-1)!}{(R-1)!\,(N-R)!} \left(\frac{y_R}{N}\right)^{R-1} \left(1 - \frac{y_R}{N}\right)^{N-R}, \qquad (3.4.2)$$

которое при $N \to \infty$ сходится к распределению вида

$$\lim_{N\to\infty} f(y_R) = \frac{1}{(R-1)!} y_R^{R-1} e^{-y_R}, \qquad (3.4.3)$$

т. е к гамма-распределению с параметром R. Например, наименьшая порядковая статистика $(R=1)$ распределена асимптотически экспоненциально.

Аналогичным образом можно получить предельное распределение для порядковых статистик, отстоящих на фиксированном месте $(N-R)$ от правого края упорядоченной статистики:

$$\lim_{N\to\infty} f(y_{N-R}) = \frac{1}{(N-R-1)!} y_{N-R}^{N-R} e^{-y_{N-R}} \qquad (3.4.4)$$

(здесь принято $y_{N-R} = N(1-x_{(N-R)})$. Чтобы получить предельное распределение крайних порядковых статистик для произвольного непрерывного распределения $F(x)$, достаточно вспомнить, что величина $z=F(x)$ распределена равномерно, и, следовательно, можно получить интересующие нас распределения, подвергнув (3.4.3) или (3.4.4) соответствующим преобразованиям.

Например, для произвольных фиксированных R и для распределений $F(x)$ экспоненциального типа Гумбель [1] получил таким методом следующее предельное распределение:

$$\lim_{N\to\infty} f(x_{(R)}) = \frac{R^R}{(R-1)!}\exp[Ry_r - Re^{-y_r}], \qquad (3.4.5)$$

где

$$y_r = \exp\left\{-\frac{R}{N}(x_{(R)}-\tilde{x})F'(\tilde{x})\right\}, \quad \tilde{x}-\text{мода } F(x).$$

Многие выдающиеся математики изучали асимптотические свойства крайних ($R=1$ или $R=N$) порядковых статистик. Наиболее полно и строго этот вопрос рассмотрел Б. В. Гнеденко [1, 2]; основные результаты в этом направлении можно изложить следующим образом (для определенности будем пока говорить только о $x_{(N)}$).

Не при любом распределении $F(x)$ исходной выборки существует устойчивое предельное распределение порядковой статистики $x_{(N)}$. Если $F(x)$ таково, что это предельное распределение существует, то последнее может быть только одного из трех типов, а именно:

$$\Lambda_1(x) = \begin{cases} 0, & x\leqslant 0 \\ \exp[-x^{-\alpha}], & x>0,\ \alpha>0; \end{cases} \qquad (3.4.6)$$

$$\Lambda_2(x) = \begin{cases} \exp[-(-x)^{\alpha}], & x\leqslant 0,\ \alpha>0, \\ 1, & x\geqslant 0; \end{cases} \qquad (3.4.7)$$

$$\Lambda_3(x) = \exp(-e^{-x}), \quad -\infty\leqslant x\leqslant\infty. \qquad (3.4.8)$$

Основная идея доказательства этого утверждения состоит в следующем. Пусть существует предельное распределение $\Lambda(x)$ для $x_{(N)}$. Разобьем выборку объема N на n подвыборок объема m; $N=mn$. Тогда $x_{(N)}$ является не только наибольшей из N величин, но и наибольшей из n наибольших в своих подвыборках величин. Так как мы предположили существование предельного распределения $\Lambda(x)$ при $N\to\infty$, то и при $m\to\infty$ распределение наибольшего в подвыборке значения должно стремиться к $\Lambda(x)$. Отсюда $\Lambda(x)$ должно обладать свойством

$$\Lambda^n(a_n x+b_n) = \Lambda(x), \qquad (3.4.9)$$

которое просто означает, что наибольшее значение в выборке объемом n из совокупности наибольших значений распределено так же, как и элементы этой совокупности — с точностью до линейного преобразования аргумента функции распределения. Решения функционального уравнения (3.4.9) и являются возможными предельными распределениями *.

Так как $\Lambda(x)$ монотонна и $0 \leqslant \Lambda(x) \leqslant 1$, то все возможные решения (3.4.9) можно разбить на три класса:

(1) $\Lambda(x)=0$ при $x \leqslant x_0$ и $\Lambda(x)>0$ при $x>x_0$,

(2) $\Lambda(x)=1$ при $x>x_0$ и $\Lambda(x)<1$ при $x \leqslant x_0$,

(3) $0<\Lambda(x)<1$ при $-\infty<x<+\infty$.

(Здесь x_0 — некоторая точка, которую без потери общности можно принять за нуль). Оказывается, что требования, предъявляемые к $\Lambda(x)$ как к функции распределения, в совокупности с необходимостью удовлетворения (3.4.9), дают единственное решение для каждого из указанных классов, которые и соответствуют распределениям (3.4.6)—(3.4.8).

Не вдаваясь в тонкости доказательств и поступясь кое-где строгостью, укажем характер необходимых и достаточных условий, которым должно удовлетворять $F(x)$, чтобы распределение $x_{(v)}$ сходилось к одному из трех асимптотических типов.

1) $\Lambda(x)=\Lambda_1(x)$ (3.4.6), если и только если

$$\lim_{x \to \infty} \frac{1-F(x)}{1-F(kx)} = k^{\alpha},$$

при любом $k>0$. (Этому требованию удовлетворяет, например, распределение Коши).

2) $\Lambda(x)=\Lambda_2(x)$ (3.4.7), если и только если

а) область возможных значений X ограничена справа величиной $x_0<\infty$, и б) для любого $k>0$

$$\lim_{x \to -0} \frac{1-F(kx+x_0)}{1-F(x+x_0)} = k^{\alpha}.$$

3) Условие для $\Lambda(x)=\Lambda_3(x)$ (3.4.8) является несколько более сложным, но по существу требует, чтобы $1-F(x)$ убывало с ростом x не медленнее, чем e^{-x}. Во всяком случае, этому условию удовлетворяют все распределения экспоненциального типа (нормальное, экспоненциальное, логнормальное, логистическое, релеевское, гамма-распределение, и т. п.), т. е.

* История здесь такова (см. Сархан и Гринберг [1], гл 7). В 1927 г. Фреше нашел одно решение уравнения (3.4.9). Фишер и Типпет в 1928 г. нашли то же и еще два решения. И, наконец, Гнеденко в 1941 г. показал, что существуют только эти три решения, и нашел необходимые и достаточные условия для существования каждого из них.

весьма часто встречающиеся распределения, благодаря чему $\Lambda_3(x)$ изучалось детальнее остальных.

До сих пор мы говорили лишь о наибольшей порядковой статистике, $x_{(N)}$. Конечно, подобные же результаты справедливы и для $x_{(1)}$: соответствующие три типа асимптотических распределений таковы:

$$\Lambda_{1'}(x)=\begin{cases} 1-\exp[-(-x)^{-\alpha}], & x\leqslant 0,\ \alpha>0,\\ 1, & x>0; \end{cases}$$

$$\Lambda_{2'}(x)=\begin{cases} 0, & x\leqslant 0\\ 1-\exp[-x^{\alpha}], & x>0,\ \alpha>0; \end{cases}$$

$$\Lambda_{3'}(x)=1-\exp(-e^{x}),\quad -\infty<x<\infty.$$

(Естественно, в условии существования $\Lambda_2(x)$ фигурирует теперь ограниченность X слева, а в условиях для Λ_1 и Λ_2 рассматриваются пределы при $x\to-\infty$). Следует иметь в виду, что $x_{(1)}$ и $x_{(N)}$ из одного и того же распределения $F(x)$ могут обладать асимптотическими распределениями как одного, так и разных типов, так как решающим является поведение $F(x)$ в разных областях значений X

§ 3.5. АСИМПТОТИЧЕСКИЕ СВОЙСТВА ВЫБОРОЧНЫХ КВАНТИЛЕЙ

Рассмотрим теперь асимптотическое поведение порядковых статистик, занимающих «средние места» в упорядоченной статистике Под этим мы будем понимать, что нас интересует асимптотическое поведение порядковой статистики при условии, что при любых N она разделяет упорядоченную статистику на две части в заданной пропорции. Для этого нам придется отказаться от фиксации ее номера R и рассмотреть поведение $x_{\widehat{R}}$ при возрастающих R с ростом N так, что $R/N=p_{\geq}=p+O(1/N)$. Ограничимся пока случаем равномерного распределения величины X и введем последовательность случайных величин $v_R=(x_{(Np_N)}-p)\sqrt{N};\ N=m,\ m+1,\ \ldots;$ $m>Np_N$ Из (3 4 1) следует, что

$$f(v_R)=\frac{N!}{(R-1)!\,(N-R)!\,\sqrt{N}}\left(p+\frac{v_R}{\sqrt{N}}\right)^{R-1}\cdot\left(1-p-\frac{v_R}{\sqrt{N}}\right)^{N-R}=\frac{N!\,(p-v_R/\sqrt{N})^{-1}}{\sqrt{N}\,(Np_N-1)!\,[(1-p_N)N]!}\,[p^{p_N}(1-p)^{1-p_N}]^{N}\times$$

$$\times\left[\left(1+\frac{v_p}{p\sqrt{N}}\right)^{p_N}\cdot\left(1-\frac{v_R}{(1-p)\sqrt{N}}\right)^{1-p_N}\right]^{N}. \qquad (3\,5\,1)$$

Так как $p_N = p + O\left(\frac{1}{N}\right)$, то

$$\left(1+\frac{v_R}{p\sqrt{N}}\right)^{p_N}\left(1-\frac{v_R}{(1-p)\sqrt{N}}\right)^{1-p_N} = \left(1-\frac{v_R^2}{2Np(1-p)}+\varphi(z, N)\right), \quad (3.5.2)$$

где $\varphi(z, N)$ таково, что $\lim\limits_{N\to\infty} N\varphi(z, N)=0$ для всех z. Используя формулу Стирлинга для факториалов, можно показать, что

$$\frac{N!\,[p^{p_N}(1-p)^{1-p_N}]^N}{\sqrt{N}\,(Np_N-1)!\,[(1-p)N]!} = \sqrt{\frac{p}{2\pi(1-p)}}+O\left(\frac{1}{N}\right). \quad (3.5.3)$$

Подставляя (3.5.2) и (3.5.3) в (3.5.1), имеем

$$f(v_R)=\left(p+\frac{v_R}{\sqrt{N}}\right)^{-1}\left(\sqrt{\frac{p}{2\pi(1-p)}}+O\left(\frac{1}{N}\right)\right)\left(1-\frac{v_R^2}{2Np(1-p)}+\varphi(z, N)\right)^N. \quad (3.5.4)$$

Теперь видно, что (3.5.4) при $N\to\infty$ равномерно сходится к функции

$$\lim_{N\to\infty} f(v_R)=\frac{1}{\sqrt{2\pi p(1-p)}}\exp\left(-\frac{v_R^2}{2p(1-p)}\right), \quad (3.5.5)$$

т. е к нормальному распределению

Переходя теперь к случаю произвольного распределения $F(x)$, получим, что асимптотическая функция распределения порядковой статистики $x_{(Np_N)}$ выразится как

$$P(x_{(Np_N)}\leqslant z)\cong\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{T(z)}\exp\left(-\frac{t^2}{2}\right)dt, \quad (3.5.6)$$

где

$$T(z)=\frac{F(z)-p}{\sqrt{p(1-p)}}\cdot\sqrt{N}.$$

Из (3.5.6) вытекает, что при $N\to\infty$ $F(x_{(Np_N)})$ сходится по вероятности к p. Следовательно, если для $F(x)$ существует единственный квантиль x_p, то $x_{(Np_N)}$ сходится к нему по вероятности. В силу этого оказывается, что в случае существования в окрестности квантиля x_p плотности вероятностей $p(x)$ $x_{(Np_N)}$ распределено асимптотически нормально со сред-

ним x_p и дисперсией $p(1-p)/Np^2(x_p)$. Например, при $p=0{,}5$ $x_{(Np/N)}$ является оценкой медианы, и если X распределено нормально с параметрами a и σ^2, то эта оценка асимптотически нормальна с параметрами a и $\pi\sigma^2/2N$.

Факт сходимости порядковой статистики к квантилю соответствующего уровня является одним из важнейших непараметрических фактов; обсудим его поэтому подробнее. Среднее значение величины $F(x_{(R)})$ равно математическому ожиданию R-й порядковой статистики $y_{(R)}$ из равномерного в [0, 1] распределения:

$$EF(x_{(R)}) = \int_0^1 y_{(R)} f(y_{(R)})\, dy_{(R)} = \frac{R}{N+1}, \qquad (3.5.7)$$

где усреднение ведется по распределению f (3.4.1). С другой стороны, только что было показано, что $F(x_{(R)})$ не только в среднем равно $R/(N+1)$ при любом N, но сходится к $R/(N+1)$ по вероятности при $N\to\infty$. Поэтому можно записать приближенное равенство

$$F(x_{(R)}) \cong \frac{R}{N+1}, \qquad (3.5.8)$$

точность которого возрастает при $N\to\infty$ и $R/(N+1)\to p=$ $=\text{const}$. Далее, если $F(x)$ непрерывна и строго монотонна, а N достаточно велико, то соотношение (3.5.8) можно обратить, что даст новое приближенное равенство:

$$x_{(R)} \cong F^{-1}\left(\frac{R}{N+1}\right). \qquad (3.5.9)$$

Соотношения (3.5.8) и (3.5.9) представляют большой интерес для непараметрической статистики, поскольку они справедливы для весьма широкого класса функций F. Так как $x_{(R)}$, R и N известны в данном эксперименте, то, например, (3.5.8) можно использовать для оценки $F(x)$, а (3.5.9) — для точечной оценки квантиля уровня $R/(N+1)$. Это, конечно, не единственно возможные использования данного непараметрического факта. В связи с этим следует упомянуть, что в некоторых случаях можно попытаться улучшить точность приближения, в особенности, формулы (3.5.9). Всякое такое уточнение, естественно, возможно лишь при дополнительной информации о распределении $F(x)$. Например, для нормального распределения (3.5.9) может быть заменено на более точное соотношение (см. Сархан и Гринберг [1])

$$x_{(R)} \simeq \Phi^{-1}\left(\frac{R-\frac{3}{8}}{N+1-\frac{3}{4}}\right).$$

Другой путь уточнения (3.5.9) состоит в разложении правой части точного равенства $x_{(R)}=F^{-1}(y_{(R)})$ в ряд по $y_{(R)}$ около точки $R/(N+1)$ и учета членов более высоких порядков, что возможно (подчеркнем это еще раз!) опять-таки лишь при известности функции F^{-1}.

В заключение параграфа приведем без доказательства (которое можно найти у Мостеллера [1]) теорему, характеризующую асимптотическое поведение совокупности выборочных квантилей:

Если $p(x)$ дифференцируема в окрестности квантилей x_{λ_i} и $p(x_{\lambda_i})\neq 0$ $(i=1, 2, \ldots, k)$, то совместное распределение $x_{(R_1)}, x_{(R_2)}, \ldots, x_{(R_k)}$ (где $x_{(R_i)}$ — выборочный квантиль уровня λ_i) является асимптотически нормальным со средними $x_{\lambda_1}, x_{\lambda_2}, \ldots, x_{\lambda_k}$ и ковариациями

$$\operatorname{cov}(x_{(R_i)}, x_{(R_j)}) = \frac{\lambda_i(1-\lambda_j)}{Np(x_{\lambda_i})p(x_{\lambda_i})}, \quad i<j.$$

ГЛАВА IV

СТАТИСТИЧЕСКИЕ СВОЙСТВА ВЫБОРОЧНЫХ ИНТЕРВАЛОВ И БЛОКОВ

§ 4.1. ВЫБОРОЧНЫЕ БЛОКИ И ИХ ПОКРЫТИЯ

Пусть $x_{(0)}, x_{(1)}, \ldots, x_{(N)}, x_{(N+1)}$ — дополненная упорядоченная статистика (см. § 3.1) одномерной выборки. Интервалы между соседними порядковыми статистиками $(x_{(0)}, x_{(1)})$, $(x_{(1)}, x_{(2)}), \ldots, (x_{(N)}, x_{(N+1)})$ называются выборочными интервалами (или промежутками); длины выборочных интервалов будем обозначать символом $\Delta_r = x_{(R)}-x_{(R-1)}$, $r=1, \ldots, N+1$. Вероятностные меры выборочных интервалов, определяемые соответственно как $u_r = F(x_{(R)})-F(x_{(R-1)})$, $r=1, 2, \ldots, N+1$, называются покрытиями (или долями). Очевидно, $\sum_{r=1}^{N+1} u_r = 1$; поэтому покрытия, рассматриваемые как случайные величины, не являются независимыми.

Аналогичные понятия можно ввести и для многомерных случайных величин. Как мы уже говорили в § 3.1, упорядочивание многомерных выборочных значений производится с помощью упорядочивающих функций $h(x_1, \ldots, x_m) = w$. Совокупности порядковых статистик $w_{(1)}, \ldots, w_{(N)}$ соответствует совокупность гиперповерхностей $h(x_1, \ldots, x_m) = w_{(R)}$, которые разбивают все пространство на $(N+1)$ подпространств $B_1^{(m)}$, $B_2^{(m)}, \ldots, B_{N+1}^{(m)}$, называемых выборочными блоками. (Верхний индекс указывает на размерность блока; выборочные интервалы есть выборочные блоки размерности единица). Таким образом, выборочным блокам одномерной величины W поставлены в соответствие m-мерные выборочные блоки величины $(X_1 \cap X_2 \cap \ldots \cap X_m)$. Покрытия $\{u_r\}$ блоков одномерной величины W одновременно являются покрытиями m-мерных блоков $\{B_r^{(m)}\}$ исходной многомерной величины.

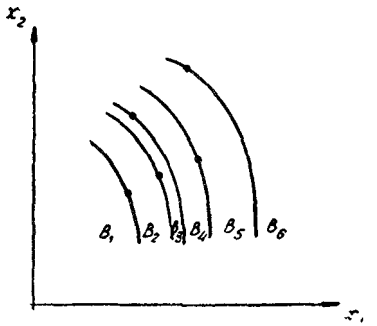

Рис 4 1.1

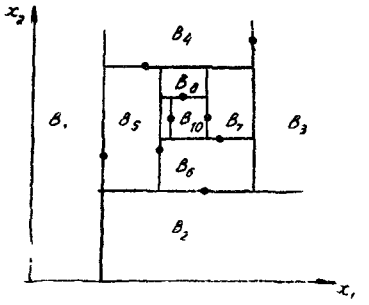

Рис 4 1 2

Обсудим несколько подробнее вопрос об упорядочивающих функциях $h(x_1, \ldots, x_m)$; для наглядности возьмем двумерный случай. Как уже упоминалось, единственное ограничение на функции $h(x_1, x_2)$ состоит в том, чтобы $F(w)$ было непрерывным; *в остальном эти функции произвольны.* Поэтому можно предложить несколько различных способов упорядочивания двумерной выборки.

Первый, очевидный, способ состоит в задании некоторой единственной аналитической функции. Тогда плоскость x_1Ox_2 разбивается на блоки $B_r^{2)}$ кривыми $h(x_1, x_2) = w_{(R)}$, проходящими через соответствующие выборочные точки $(x_{1(R)}, x_{2(R)})$ (см. рис. 4.1.1).

Другой способ, предложенный Вальдом [1] и обобщенный Тьюки [1], по сути, состоит в последовательном использовании нескольких упорядочивающих функций, $\{h_l(x_1, x_2)\}$, $l = 1, 2, \ldots, N$. Первая кривая $w_{(1)} = h_1(x_1, x_2)$ делит плоскость

x_1Ox_2 на две части, $B_1^{(2)}$ (такую, что $h_1(x_1, x_2) < w_{(1)}$, если $(x_1, x_2) \in B_1^{(2)}$) и $\overline{B}_1^{(2)}$). Вторая кривая таким же способом делит область $\overline{B}_1^{(2)}$ на $B_2^{(2)}$ и $\overline{B}_2^{(2)}$ и т. д. На рис. 4.1.2 показан пример разбиения на блоки, когда кривые $h_l(x_1, x_2)$ являются лучами и отрезками прямых.

Фрейзер [1] предложил еще несколько способов упорядочивания многомерных выборок — с помощью последовательности кривых первого и второго порядка.

Как видим, упорядочивание многомерной выборки, а с ним и разбиение выборочного пространства на блоки не являются однозначными. Хотя, как это будет показано ниже, определенные статистические свойства выборочных блоков не зависят от выбора упорядочивающих функций, в конкретных статистических задачах некоторые упорядочивающие функции могут оказаться предпочтительнее других.

§ 4.2. РАСПРЕДЕЛЕНИЕ ПОКРЫТИЙ И ИХ СУММ

Согласно определению, приведенному в § 4.1, покрытия (т. е. вероятностные меры интервалов между соседними порядковыми статистиками) есть функции $\{u_r = F(x_{(R)}) - F(x_{(R-1)})\}$, $r = 1, 2, \ldots, N+1$, и как таковые являются случайными переменными. Свойства покрытий очень интересны и полезны для построения непараметрических процедур.

Рассмотрим сначала случайные величины $y_{(R)} = F(x_{(R)})$. Очевидно, $y_{(R)} = \sum_{i=1}^{R} u_i$. С другой стороны, все $\{y_{(R)}\}$ статистически независимы и равномерно распределены в интервале $[0, 1]$. Поскольку $(y_{(1)}, y_{(2)}, \ldots, y_{(N)})$ является упорядоченной статистикой выборки из равномерного распределения, то любые интересующие нас распределения величин $\{y_{(R)}\}$ могут быть получены как частные случаи формул предыдущей главы Так, согласно (3.2.2) имеем:

$$f(y_{(1)}, y_{(2)}, \ldots, y_{(N)}) = N!, \quad 0 \leqslant y_{(1)} \leqslant \ldots \leqslant y_{(N)} \leqslant 1. \qquad (4.2.1)$$

Из формулы (3.2.10) следует:

$$f(y_{(k_1)}, y_{(k_1+k_2)}, \ldots, y_{(k_1+k_2+\ +k_s)}) =$$

$$= \frac{N!}{(k_1-1)!\,(k_2-1)!\ldots(k_s-1)!\,(N-k_1-k_2-\ldots-k_s)!} \times$$

$$\times y_{(k)}^{k_1-1} \cdot (y_{(k_1-k_2)} - y_{(k_1)})^{k_2-1} \ldots [1 - y_{(k_1+k_2+\ +k_s)}]^{-N-k-\ -k_s}. \qquad (4.2.2)$$

С помощью этого распределения легко найти распределение сумм любых покрытий. Найдем, например, распределение

размаха выборки, который определяется как разность наибольшей и наименьшей порядковых статистик,

$$v_{1N}=y_{(N)}-y_{(1)},$$

т. е. является суммой покрытий всех интервалов между $x_{(1)}$ и $x_{(N)}$. Из (4.2.2) при $k_1=1$, $k_s=N-1$, $s=2$ имеем:

$$f(y_{(1)}, y_{(N)})=\frac{N!}{(N-2)!}(y_{(N)}-y_{(1)})^{N-2}.$$

От переменных $(y_{(1)}, y_{(N)})$ перейдем к переменным $(y_{(1)}, v_{1N})$. Якобиан перехода равен 1, следовательно

$$f(y_{(1)}, v_{1N})=N(N-1)v_{1N}^{N-2},\ 0\leqslant y_{(1)}\leqslant(1-v_{1N})\leqslant 1.$$

Интегрируя $f(y_{(1)}, v_{1N})$ по $y_{(1)}$, получаем:

$$f(v_{1N})=N(N-1)v_{1N}^{N-2}(1-v_{1N}),\ 0\leqslant v_{1N}\leqslant 1.$$

Более общо, сумма покрытий между любыми $x_{(R)}$ и $x_{(S)}$ $(R<S)$ распределена согласно (Уилкс [1])

$$f(v_{RS})=\frac{N!}{(S-R-1)!\,(N-S+R)!}v_{RS}^{S-R-1}(1-v_{RS})^{N-S+R},$$
$$0\leqslant v_{RS}\leqslant 1. \qquad (4.2.3)$$

§ 4.3. СТАТИСТИЧЕСКАЯ ЭКВИВАЛЕНТНОСТЬ ВЫБОРОЧНЫХ БЛОКОВ

Замечательным свойством выборочных блоков является их статистическая эквивалентность. Это свойство выражается в том, что ни распределение любого из покрытий, ни распределения любой совокупности покрытий не зависят от их номеров.

В самом деле, из (4.2.3) при $S=R+1$ получаем для любого R:

$$f(u_r)=N(1-u_r)^{N-1},\ 0\leqslant u_r\leqslant 1. \qquad (4.3.1)$$

Далее, поскольку якобиан перехода от $\{y_{(R)}\}$ к $\{u_r\}$ равен единице, совместное распределение всех покрытий, согласно (4.2.1), равно

$$f(u_1, u_2, \ldots, u_N)=N!,\ \left(\sum_{i=1}^{N}u_i\leqslant 1\right), \qquad (4.3.2)$$

откуда можно найти любые интересующие нас распределения. Например, для любых R и S

$$f(u_r, u_s)=N(N-1)(1-u_r-u_s)^{N-2}. \qquad (4.3.3)$$

Более общо, любые k покрытий ($k \leqslant N$), $u_{1'}$, $u_{2'}$, ..., $u_{k'}$ (с любыми исходными номерами, для чего и введены штрихованные индексы) распределены согласно плотности

$$f(u_{1'}, u_{2'}, \ldots, u_{k'}) = \frac{N!}{(N-k)!}(1-u_{1'}-u_{2'}-\ldots-u_{k'})^{N-k} \qquad (4.3.4)$$

в области, определенной соотношениями $\sum_{i=1'}^{k'} u_i \leqslant 1$ и $\{u_i \geqslant 0\}$. Как видим, распределения (4.3.1), (4.3.3), (4.3.4) не зависят от конкретных номеров покрытий, о которых идет речь. Это и имеют в виду, когда говорят, что порядковые статистики $\{x_{(R)}\}$ делят область возможных значений случайной величины на статистически эквивалентные блоки.

Прямым следствием статистической эквивалентности блоков является независимость числовых характеристик распределений их покрытий от их номеров.

Так, среднее значение любого из покрытий равно

$$E(u_r) = \int_0^1 N(1-u_r)^{N-1} du_r = \frac{1}{N+1}, \qquad (4.3.5)$$

т. е. N упорядоченных значений случайной величины X делят площадь под неизвестной плотностью $p(x)$ на $N+1$ в среднем равных частей — интересный непараметрический факт, который может быть использован в дальнейшем.

Далее, с помощью приведенных выше распределений легко получить:

$$E(u_r^2) = \frac{2}{(N+1)(N+2)}, \qquad (4.3.6)$$

$$E\left[\left(u_r - \frac{1}{N+1}\right)^2\right] = \frac{N}{(N+1)^2(N+2)}, \qquad (4.3.7)$$

$$E(u_k \cdot u_r) = \frac{1}{(N+1)(N+2)}, \qquad (4.3.8)$$

$$E\left[\left(u_r - \frac{1}{N+1}\right)\left(u_s - \frac{1}{N+1}\right)\right] = \frac{-1}{(N+1)^2(N+2)}. \qquad (4.3.9)$$

С помощью (4.3.7) и (4.3.9) легко вычислить коэффициент корреляции между u_r и u_s, который оказывается равным $(-1)/N$. Эту ослабевающую с увеличением объема выборки корреляционную связь и имеют в виду, когда говорят о том, что при больших объемах выборки блоки становятся «почти независимыми».

Еще более интересным является то, что статистической эквивалентностью обладают и блоки многомерных случайных величин. Многомерные выборочные блоки получаются путем использования некоторой последовательности упорядочивающих функций; в результате этого устанавливается отображение блоков многомерной величины на блоки некоторой одномерной. А так как блоки *любой* непрерывной одномерной величины, каково бы ни было ее распределение, обладают статистической эквивалентностью, то этим же свойством обладают и блоки соответствующей многомерной величины. Замечательно здесь то, что это свойство сохраняется несмотря на довольно большой произвол при выборе упорядочивающих функций.

§ 4.4. ОБЩИЕ СВОЙСТВА ВЫБОРОЧНЫХ ИНТЕРВАЛОВ

Иногда представляют интерес не покрытия, а сами выборочные интервалы между соседними порядковыми статистиками. Естественно, что свойства выборочных интервалов существенно зависят от исходного распределения $p(x)$ выборки*; однако оказывается, что и для произвольных $p(x)$ выборочные интервалы обладают некоторыми общими свойствами.

Совместное распределение порядковых статистик дается формулой (3.2.2). Поскольку выборочные интервалы Δ_i являются линейными функциями порядковых статистик ($\Delta_i = x_{(i)} - x_{(i-1)}$. $x_{(k)} = x_{(1)} + \Delta_1 + \Delta_2 + \ldots + \Delta_k$), то из (3.2.2) легко получить совместное распределение выборочных интервалов

$$f(\Delta_2, \Delta_3, \ldots, \Delta_N) = N! \int_{-\infty}^{\infty} \prod_{i=2}^{N} p(x+\Delta_2+\ldots+\Delta_i)\, dx \qquad (4.4.1)$$

(f отлично от нуля в области $\Delta_i > 0$, $2 \leqslant i \leqslant N$). Распределение i-го интервала может быть получено аналогичным путем из (3.2.10):

$$f(\Delta_i) = \frac{N!}{(i-2)!\,(N-i)!} \int_{-\infty}^{\infty} [F(x)]^{i-2} [1 - F(x+\Delta_i)]^{N-i} p(x)\, p(x+\Delta_i)\, dx \qquad (4.4.2)$$

Как видим, получить явные выражения при конкретных распределениях $p(x)$ довольно сложно: необходимо вычислять соответствующие интегралы, которые обычно не сведутся к элементарным функциям.

---

* Напомним кстати, что покрытия являются выборочными интервалами для равномерного в [0,1] распределения

Кроме случая равномерного распределения (см. § 4.2) известен еще один пример, когда вычисления могут быть доведены до конца аналитически, а именно — случай экспоненциального распределения $p(x)=\lambda \exp(-\lambda x)$, $x \geqslant 0$ (Пайк [1]) При этом

$$f(\Delta_1, \Delta_2, ., \Delta_N)=N!\lambda^N \prod_{i=1}^{N} \exp[-\lambda(\Delta_1+...+\Delta_i)]=$$

$$=N!\lambda^N \exp\left[-\lambda \sum_{i=1}^{N}(N-i+1)\Delta_i\right]=$$

$$=\prod_{i=1}^{N}\lambda(N-i+1)\exp[-\lambda(N-i+1)\Delta_i]. \qquad (4.4\,3)$$

Интересно, что в этом случае выборочные интервалы оказываются статистически независимыми (так как совместное распределение факторизовалось) экспоненциально распределенными величинами с параметрами соответственно λN, $\lambda(N-1)$, ..., λ. Если теперь ввести нормализованные выборочные интервалы $\delta_i=\lambda(N-i+1)\Delta_i$, то они оказываются независимыми и одинаково (экспоненциально) распределенными с единичным средним значением.

Значение этого, казалось бы, частного, хотя и любопытного результата резко возрастает, если учесть, что выборка из любого непрерывного распределения $F(x)$ может быть преобразована в выборку из экспоненциального распределения с помощью соотношения $Y_i=-\log(1-F(X_i))$. С другой стороны, порядковые статистики и выборочные интервалы произвольного непрерывного распределения, в свою очередь, могут быть выражены через выборочные значения из экспоненциального распределения (Сухатме [1]). Использование этих фактов позволяет значительно расширить сведения об общих свойствах выборочных интервалов и их покрытий. Покажем это на ряде примеров.

1. Покрытия как экспоненциальные случайные величины, отнесенные к их сумме.

Пусть Y_1, Y_2, ..., Y_{N+1} — независимые экспоненциально распределенные случайные величины с единичным средним и $S=Y_1+Y_2+...+Y_{N+1}$. Тогда величины $u_i=Y_i/S$, $(1 \leqslant i \leqslant N+1)$ распределены так же, как выборочные интервалы для выборки объема N из равномерного в [0, 1] распределения (т. е. так же, как покрытия).

В самом деле, так как совместное распределение Y-ов известно

$$p(Y_1, ., Y_{N+1})=e^{-(Y_1+Y_N)}, \qquad (4.4.4)$$

то

$$p(Y_1, ..., Y_N, S)=e^{-S}, \qquad (4.4\,5)$$

откуда

$$p(u_1,\ldots, u_N, S)=e^{-S}\cdot S^N, \quad \left(0\leqslant \sum_1^N u_1\leqslant 1\right). \qquad (4.4.6)$$

Интегрируя (4.4.6) по S, получаем:

$$p(u_1,\ldots, u_N)=N!, \qquad (4.4.7)$$

что совпадает с распределением покрытий (4.3.2).

2. Покрытия как функции экспоненциально распределенных случайных величин

Порядковые статистики выборки из равномерного в [0, 1] распределения (и, следовательно, покрытия) могут быть записаны в форме некоторых функций от порядковых статистик выборки из экспоненциального распределения. Пусть $Y_1, Y_2, \ldots, Y_N$ — независимые экспоненциальные случайные величины с единичным средним. Введем случайные величины Z_i и U_i:

$$Z_i=\sum_{j=1}^{i}(N-j+1)^{-1}Y_j, \quad U_i=1-\exp(-Z_i). \qquad (4.4.8)$$

Из (4.4.3) и последующего обсуждения следует, что $(Y_1/N, Y_2/(N-1), \ldots Y_N)$ распределены так же, как выборочные интервалы выборки из экспоненциального распределения с единичным средним, а значит $(Z_1, Z_2, \ldots, Z_N)$ можно рассматривать как порядковые статистики этой выборки. Тогда U_i (в силу того, что $1-\exp(-Z)$ является интегральной функцией распределения) распределены так же, как порядковые статистики выборки из равномерного в [0, 1] распределения. Следовательно, i-е покрытие можно записать в виде:

$$u_i=U_i-U_{i-1}=\exp(-Z_{i-1})-\exp(-Z_i)=$$
$$=(1-U_{i-1})[1-\exp[-Y_i'(N-i+1)]. \qquad (4.4.9)$$

С помощью представлений (4.4.8) и (4.4.9) можно сравнительно просто получить ряд важных теоретических результатов. Например, можно доказать, что отношения соседних порядковых статистик для равномерного распределения статистически независимы. Действительно, из (4.4.8) следует, что

$$\frac{U_{N-i+1}}{U_{N-i}}=\exp\left(-\frac{Y_{N-i+1}}{i}\right), \qquad (4.4.10)$$

а так как $\{Y_i\}$ независимы, то и соответствующие отношения U_1/U_2, U_2/U_3, ..., U_{N-1}/U_N, U_N тоже независимы. Другой пример использования указанных соотношений (Реньи [1]) состоит в том, что при рассмотрении асимптотических

проблем величина $\left(U_i-\frac{i}{N+1}\right)$ может быть заменена величиной

$$W_i=\left(1-\frac{i}{N}\right)\sum_{j=1}^{i}\frac{Y_i-1}{N-j+1}. \qquad (4.4.11)$$

Преимущество использования W_i вместо $\left(U_i-\frac{i}{N+1}\right)$ состоит в том, что W_i является суммой независимых случайных величин (с нулевым средним), что позволяет использовать классические предельные теоремы. (4.4.11) является аппроксимацией, качество которой повышается с ростом объема выборки. Действительно, $U_i=1-\exp(-Z_i)\approx Z_i$, а $1-\frac{i}{N}\approx$ $\approx\exp[-\sum_{j=1}^{i}(N-j+1)^{-1}]$; откуда и получается (4.4.11). Можно доказать (Поршан и Пайк [1]), что для всех $i=1, 2, ..., N$

$$E\left|U_i-\frac{i}{N+1}-W_i\right|<CN^{-1}, \qquad (4.4.12)$$

где C — константа, не зависящая от i и N. Соотношение (4.4.12) доказывается тем, что

$$E\left[U_i-\frac{i}{N+1}-W_i\right]^2=\operatorname{var}U_i+EW_i^2+2E(W_i\exp(-Z_i))$$

ограничено сверху величиной C^2N^{-2}, что проверяется непосредственной оценкой величин в правой части последнего равенства. Соотношение (4.4.12) не только обосновывает возможность использования W_i в качестве аппроксимации U_i но и дает возможность оценить ошибку такой аппроксимации.

3. Выборочные интервалы произвольного распределения как функции интервалов экспоненциального распределения.

Пусть $T_i=F^{-1}(U_i)$, где F — произвольная непрерывная функция распределения. Так как U^i является порядковой статистикой равномерного распределения, то T_i оказывается порядковой статистикой распределения F. Следовательно, выборочные интервалы для этого распределения могут быть записаны в виде

$$\Delta_i=F^{-1}(U_i)-F^{-1}(U_{i-1}),\ 2\leqslant i\leqslant N. \qquad (4.4.13)$$

Далее, если $H(z)$ — некоторая другая функция распределения, то (4.4.13) можно записать в виде:

$$\Delta_i = F^{-1}(H(Z_i)) - F^{-1}(H(Z_{i-1})), \qquad (4.4.14)$$

где Z_i — порядковая статистика выборки из распределения H. Удобно воспользоваться экспоненциальным распределением в качестве H, так как оно дает статистически независимые порядковые статистики и интервалы. Предположив существование производной $k(Z)$ от функции $F^{-1}(H(Z))$ и воспользовавшись теоремой о среднем, имеем:

$$\Delta_i = (Z_i - Z_{i-1}) \cdot k(a_i), \qquad (4.4.15)$$

где $Z_{i-1} \leqslant a_i \leqslant Z_i$, или в явном виде:

$$\Delta_i = (Z_i - Z_{i-1}) e^{-a_i} f[F^{-1}(H(a_i))], \qquad (4.4.16)$$

где $f(x)$ — плотность, соответствующая распределению F. Так как $(Z_i - Z_{i-1})$ равно $Y_i/(N-i+1)$, то выборочный интервал для произвольного распределения окончательно выражается через экспоненциально распределенную случайную величину:

$$\Delta_i = \frac{Y_i}{N-i+1} \cdot r(A_i), \quad (2 \leqslant i \leqslant N), \qquad (4.4.17)$$

где $r(u) = (1-u)/f(F^{-1}(u))$ и A_i лежит между порядковыми статистиками U_i и U_{i-1} равномерного распределения. Это соотношение может с успехом использоваться при изучении предельных распределений функций выборочных интервалов (Пайк [1]).

§ 4.5. АСИМПТОТИЧЕСКИЕ СВОЙСТВА ВЫБОРОЧНЫХ ИНТЕРВАЛОВ

Поскольку длины выборочных интервалов при неограниченном увеличении объема выборки неограниченно убывают, удобно ввести в рассмотрение величины $\delta_i = N\Delta_i$ и найти их предельные распределения.

Проще всего находится предельное распределение в экспоненциальном случае. Так как δ_i при этом имеет экспоненциальное распределение с параметром $\lambda(N-i+1)/N$, то, очевидно, что при $N \to \infty$ и $i/N \to u$, δ_i в пределе оказывается экспоненциально распределенной величиной с параметром $\lambda(1-u)$. Независимость δ_i и δ_j сохраняется и, следовательно, предельная совместная функция распределения величин δ_i и δ_j (при $j/N \to v$) будет равна

$$\lim_{N \to \infty} F(\delta_i, \delta_j) = [1 - e^{-\lambda(1-u)\delta_i}][1 - e^{-\lambda(1-v)\delta_j}]. \qquad (4.5.1)$$

Пожалуй, наиболее интересным асимптотическим свойством выборочных интервалов является то, что они оказыва-

ются асимптотически экспоненциальными и независимыми при любой плотности $g(x)$. Этот результат может быть сформулирован в виде следующей теоремы.

Теорема (Пайк [1]). Если для заданных u и v $(0<u<v<1)$ функции $G^{-1}(u)$ и $G^{-1}(v)$ однозначно определены, то при $i/N \to u$ и $j/N \to v$

$$\lim_{N\to\infty} F(\delta_i, \delta_j) = [1 - e^{g(G^{-1}(u))\delta_i}][1 - e^{-g(G^{-1}(v))\delta_j}]. \quad (4.5.2)$$

Доказательство теоремы основывается на сходимости по вероятности порядковых статистик к соответствующим квантилям.

§ 4.6. О МОМЕНТАХ ВЫБОРОЧНЫХ ИНТЕРВАЛОВ

Хотя вся информация о статистических свойствах выборочных интервалов и их сумм содержится в соответствующих распределениях, определенный интерес представляет отдельное рассмотрение моментов этих величин и величин, им обратных.

Пусть Δ_{KR} — интервал между K-й и R-й порядковыми статистиками; тогда $E(\Delta_{KR}^{\alpha})$ является при $\alpha>0$ начальным моментом суммы $(R-K)$ выборочных интервалов, а при $\alpha<0$ — начальным моментом величины, обратной этой сумме.

Рассмотрим сначала моменты сумм покрытий (т. е. моменты интервалов между порядковыми статистиками выборки из равномерного распределения).

$$E(\Delta_{KR}^{\alpha}) = \frac{N!}{(R-K-1)!\,(N-R+K)!}\int_0^1 \Delta_{KR}^{R-K-1+\alpha}(1-$$

$$-\Delta_{KR})^{N-R+K}\,d\Delta_{KR} = \frac{N!}{(R-K-1)!\,(N-R+K)!}\cdot$$

$$\cdot B(R-K+\alpha;\ N-R+K+1) = \frac{N!\,(R-K+\alpha-1)!}{(R-K-1)!\,(N+\alpha)!}. \quad (4.6.1)$$

Так как в некоторых статистических процедурах используются величины, обратные выборочным интервалам, то целесообразно подчеркнуть, что не у всех таких величин существуют моменты. Считая условие $(N+\alpha)\geqslant 0$ несущественным, обратим внимание на требование $R-K+\alpha-1\geqslant 0$, эквивалентное конечности момента α-го порядка (см. (4.6.1)). Отсюда следует, что для обратных сумм покрытий начальный момент α-го порядка существует только в том случае, если число слагаемых в сумме по крайней мере на единицу превышает α.

Для экспоненциального исходного распределения моменты сумм выборочных интервалов получатся путем усреднения этих сумм по распределению (4.4.3). Так как в этом случае выборочные интервалы независимы, то условие существования моментов сумм интервалов совпадает с условием существования моментов одного интервала. Легко показать, что для $\alpha>0$

$$E\Delta_R^{\alpha}=\alpha!\,[\lambda(N-R+1)]^{-\alpha}. \qquad (4.6.2)$$

Вычисление моментов величин, обратных суммам выборочных интервалов для экспоненциального распределения, приводит к расходящимся интегралам.

ГЛАВА V

СТАТИСТИЧЕСКИЕ СВОЙСТВА РАНГОВ

§ 5.1. РАНГИ. УСЛОВИЯ ИНФОРМАТИВНОСТИ РАНГОВ

Определим сначала некоторые понятия. Р а н г о м данного значения x_l из выборки заданного объема N называется число R_l выборочных значений, не превышающих x_l, т. е. удовлетворяющих условию $x_k\leqslant x_l$, $k=1, 2, \ldots, N$. (Будем пока говорить лишь о непрерывных случайных величинах; дискретный случай обсудим в § 5.5). Если ввести ф у н к ц и ю с р а в н е н и я $C(t)$,

$$C(t)=\begin{cases}1, & t\geqslant 0,\\ 0, & t<0,\end{cases} \qquad (5.1.1)$$

то R_l можно выразить формулой

$$R_l=\sum_{k=1}^{N} C(x_l-x_k). \qquad (5.1.2)$$

Так как ранг R_l является функцией выборочных значений, его можно рассматривать как случайную величину с возможными значениями 1, 2, ..., N.

Совокупность рангов одномерной выборки $\{R_l\}=(R_1, R_2, \ldots, R_N)$ называется р а н г о в ы м в е к т о р о м. Ранговый вектор тоже случаен — в том смысле, что для разных выборок из одной совокупности будут получаться различные последовательности рангов. Ясно, что реализации рангового вектора являются перестановками чисел 1, 2, ..., N; число возможных реализаций рангового вектора равно $N!$

Целесообразность рассмотрения рангов основывается на следующих соображениях. Во-первых, ранговый вектор $\{R_i\}$ содержит какую-то часть информации, содержащейся в исходной выборке $\{x_i\}$. В самом деле, с помощью операции упорядочивания выборке $\{x_i\}$ ставится в однозначное соответствие пара векторов — упорядоченная статистика $\{x_{(R)}\}$ и ранговый вектор $\{R_i\}$; и, наоборот, располагая векторами $\{x_{(R)}\}$ и $\{R_i\}$, можно однозначно восстановить выборку $\{x_i\}$. Это означает, что пара $(\{x_{(R)}\}, \{R_i\})$ содержит ту же информацию, что и $\{x_i\}$, а следовательно, некоторая доля всей информации приходится и в отдельности на $\{R_i\}$. Отсюда появляется (оправдывающаяся впоследствии) надежда, что можно строить статистические процедуры *только на рангах*, не используя знание самих реализовавшихся выборочных значений. Первое, бросающееся в глаза преимущество таких процедур состоит в целочисленности рангов и вытекающей отсюда простоте технической реализации процедур. Далее оказывается, что при определенных условиях ранговый вектор содержит не просто долю, а значительную долю полезной информации, что обеспечивает некоторым ранговым процедурам высокую эффективность. Наконец, и это в особенности важно, ранговые процедуры обладают свойством непараметричности.

Все приведенные выше утверждения будут подробно обоснованы в последующих параграфах данной главы и в главе о ранговых процедурах. А пока сделаем некоторые дополнительные замечания об особенностях ранговых процедур.

Важным достоинством этих процедур является то, что они применимы и в тех случаях, когда наблюдения носят не количественный, а качественный характер, лишь бы наблюдения допускали их упорядочивание. Например, прослушав две группы скрипачей, обучавшихся каждая по своему методу, компетентное жюри может упорядочить всех исполнителей по качеству игры (т. е. присвоить им ранги), после чего с помощью подходящего теста можно сделать статистический вывод об относительных достоинствах сравниваемых методов обучения. Аналогичным способом проводится сравнение различных медикаментов комплексного действия и т. п. Естественно, представительность, случайность выборки при таких экспериментах должны быть предметом особой заботы экспериментатора (как, впрочем, и в любом статистическом испытании).

Перейдем теперь к обсуждению весьма важного вопроса об условиях, обеспечивающих информативность рангов. Главным здесь является выбор *способа ранжировки* наблюдений. В постановке всякой статистической задачи фигури-

руют распределения, чем-то различающиеся между собой: в задачах оценки параметров распределения отличаются значениями этих параметров (хотя и не обязательно только этими значениями); в задачах проверки гипотез всегда указывается различие между нулевой и альтернативной гипотезой. Ясно, что если мы хотим решать ту или иную статистическую задачу с помощью ранговой процедуры, то прежде всего необходимо обеспечить, чтобы ранги были чувствительными к различиям между распределениями, характерным для данной задачи. Выражаясь точнее, статистические свойства рангов должны известным образом изменяться при смене различаемых распределений. Это достигается выбором подходящего способа ранжировки.

Представим себе, например, что мы хотим по одной выборке проверить, является ли случайная величина X симметрично распределенной относительно нуля, и по некоторым причинам ограничиваем себя только ранговыми процедурами. Если произвести обычную ранжировку, непосредственно по формуле (5.1.2), то что бы мы потом ни дедали с полученным таким образом ранговым вектором, нашу задачу мы не решим: этот вектор не изменится, какой бы сдвиг от нуля ни имела бы исходная выборка. С другой стороны ясно, что несимметричность распределения $p(x)$ относительно нуля должна проявляться в статистическом преобладании выборочных значений одного знака. Раз уж мы решили пользоваться только ранговой процедурой, то надо выбрать такой способ ранжировки, при котором преобладание выборочных значений одного знака сказывалось бы на свойствах рангового вектора. Так мы приходим к мысли об упорядочивании модулей $|x_i|$ выборочных значений с сохранением информации о знаках x_i. Используя эту информацию, можно затем отделить ранги модулей отрицательных значений x_i от рангов положительных выборочных значений, образовав таким образом два ранговых вектора $\{R_i^+\}$ и $\{R_j^-\}$. Очевидно, при симметрии $p(x)$ эти ранговые векторы статистически эквивалентны, а при нарушении симметрии эта эквивалентность исчезает. Как конкретно использовать это различие между $\{R_i^+\}$ и $\{R_j^-\}$ — это особый вопрос, а для нас сейчас главное то, что информация о существенном для задачи различии распределений перенесена на ранги с помощью специального способа ранжировки.

Другой пример дают двувыборочные задачи на одинаковость распределений обеих выборок, $\{x_i\}$ и $\{y_j\}$. Наиболее употребительный (и единственный?) способ построения ранговых векторов, чувствительных к различию распределений X и Y, состоит в объединении выборок $\{x_i\}$ и $\{y_j\}$ в одну

смешанную выборку* $\{w_l\}$, которая упорядочивается — с сохранением признака принадлежности w_l к $\{x_i\}$ или $\{y_j\}$. С помощью этого признака все ранги $\{R_l\}$ смешанной выборки можно разбить на две группы — ранги $\{R^{lX}\}$, принадлежащие значениям $w_l \in \{x_i\}$, и ранги $\{R_l^Y\}$, соответствующие $w_l \in \{y_j\}$. Свойства полученных ранговых векторов теперь существенно зависят от того, одинаково ли распределены X и Y. В самом деле, при одинаковых распределениях X и Y их выборочные значения в смешанной выборке имеют статистическую тенденцию к равномерному перемешиванию; и, наоборот, при различии распределений будет происходить их статистическое разделение в смешанной выборке. Например, при альтернативе сдвига статистически меньшая величина будет иметь тенденцию к обладанию меньшими рангами; при различии дисперсий — значения менее разбросанной величины имеют тенденцию занимать в смешанной выборке средние места; и т. д. Такая информативность ранговых векторов опять-таки обеспечена за счет разумно выбранного способа ранжировки.

Чтобы еще раз подчеркнуть связь между способом ранжировки и типом решаемой задачи, приведем третий пример. Пусть имеется двумерная выборка $\{(x_i, y_i)\}$; требуется установить наличие или отсутствие статистической связи между компонентами X и Y выборки, и сделать это необходимо, используя только ранговые методы. В данном случае выборки $\{x_i\}$ и $\{y_i\}$ упорядочиваются каждая в отдельности с сохранением исходного порядкового номера выборочного значения. В результате двумерной выборке $\{(x_i, y_i)\}$ объема N ставится в соответствие $(2\times N)$-мерная ранговая матрица $\{(R_i^X, R_i^Y)\}$. Хотя это и не так очевидно, как в предыдущих двух примерах, но наличие или отсутствие связи между X и Y при таком способе ранжировки приводит к наличию или отсутствию связи между R^X и R^Y, которую затем и обнаруживают тем или иным образом. Впоследствии мы будем говорить об этом детальнее, а сейчас снова подчеркнем, что и здесь способ ранжировки выбран так, чтобы в рангах осталась существенная для задачи информация.

Итак, решение всякой статистической задачи ранговыми методами начинается с выбора такого способа упорядочивания, при котором изменения исследуемого фактора приводят к известному изменению свойств рангов, ранговых векторов или матриц.

---

* Разумеется, величины X и Y должны быть одной физической природы и размерности.

Следующий подготовительный шаг во всякой ранговой процедуре состоит в отборе или рассортировке полученных рангов: в первом примере нам потребовалось отделить ранги положительных величин от рангов модулей отрицательных; во втором примере — ранги x-ов от рангов y-ов; и т. п. Во всех примерах специально подчеркивалось, что при ранжировке сохраняется признак принадлежности данного выборочного значения к тому или иному подмножеству; именно по этому признаку и осуществляется последующий отбор нужных рангов.

Для того, чтобы формализовать операцию отбора и иметь возможность выражать ее аналитически, вводится специальный оператор, называемый *индикатором* и обозначаемый символом $Z_{i\alpha}$. Индекс i индикатора есть натуральное число, индекс α характеризует признак того подмножества, которое нас интересует, а сам индикатор принимает либо нулевое, либо единичное значение. Индикатор, по существу, является обобщением символа Кронекера δ_{ik}; в отличие от последнего, индексы индикатора имеют различную природу, и один из них (α) может быть вообще не числовым.

Предположим, что символ Кронекера можно применять для индикации множеств: пусть $\delta_{\beta}=1$, если множества α и β совпадают, и $\delta_{\alpha\beta}=0$, если они различны. Тогда индикатор $Z_{i\alpha}$ можно представить как произведение двух символов Кронекера:

$$Z_{i\alpha}=\delta_{ik}\,\delta_{\alpha\beta}=\begin{cases}1, & i=k \text{ и } \alpha=\beta\\ 0, & i\neq k \text{ или } \alpha\neq\beta.\end{cases}$$

Индикатор можно употреблять и как показатель степени, и как множитель. Так, примененный к рангу K_l элемента w_l смешанной выборки из второго примера индикатор дает:

$$K_l^{Z_{iX}}=\begin{cases}i, & \text{если } l=i \text{ и } w_l\in X,\\ 1 & \text{в остальных случаях,}\end{cases} \tag{5.1.3}$$

$$K_l\cdot Z_{iX}=\begin{cases}i, & \text{если } l=i \text{ и } w_l\in X\\ 0 & \text{в остальных случаях.}\end{cases} \tag{5.1.4}$$

Естественно, индикатор можно употреблять не только по отношению к рангам, но и к элементам любых множеств, состоящих из дискретных подмножеств.

§ 5.2. СВОЙСТВА РАНГОВЫХ ВЕКТОРОВ ПРИ ИНВАРИАНТНОСТИ К ПЕРЕСТАНОВКАМ

Как явствует из рассуждений предыдущего параграфа, статистические свойства рангов прежде всего определяются

двумя факторами: 1) статистическими свойствами выборки и 2) способом упорядочивания выборочных значений. В данном параграфе рассмотрим свойства ранговых векторов для случая, когда распределение выборки инвариантно по отношению к перестановкам аргументов, т. е. когда распределение симметрично относительно своих переменных (например, если выборочные значения независимы и одинаково распределены).

Сначала рассмотрим случай, когда упорядочивание производится непосредственно по алгебраическим значениям измеряемых величин (будем по-прежнему пока рассматривать непрерывный случай). Как отмечалось в § 5.1, совокупность упорядоченной статистики $\{x_{(R)}\}$ и рангового вектора $\{R_i\}$ находится во взаимно однозначном соответствии с исходной выборкой $\{x_i\}$. Следовательно, совместное распределение f векторов $\{x_{(R)}\}$ и $\{R_i\}$ совпадает с распределением выборки:

$$p(x_1,\ldots, x_N)=p(x_{(R_1)},\ldots, x_{(R_N)})=f(\{x_{(R)}\}, \{R_i\}). \quad (5.2.1)$$

При инвариантности к перестановкам (5.2.1) упрощается:

$$f(\{x_{(R)}\}, \{R_i\})=p(x_{(1)}, x_{(2)},\ldots x_{(N)}). \quad (5.2.2)$$

(Во избежание недоразумений следует помнить, что хотя в (5.2.2) ранговый вектор явно не фигурирует, он исчез при переходе от (5.2.1) только из-за предположений инвариантности).

Учитывая, что безусловное распределение рангового вектора (при соблюдении инвариантности!) является равномерным:

$$f(\{R_i\})=\frac{1}{N!}, \quad (5.2.3)$$

что безусловное распределение упорядоченной статистики дается (см. (3.2.2)) формулой:

$$f(\{x_{(R)}\})=N!\ p(x_{(1)},\ldots, x_{(N)}) \quad (5.2.4)$$

и что (5.2.2) можно, следовательно, записать как

$$f(\{x_{(R)}\}, \{R_i\})=\frac{1}{N!}\cdot N!\ p(x_{(1)},\ldots, x_{(N)})=$$
$$=f(\{x_{(R)}\})\cdot f(\{R_i\}), \quad (5.2.5)$$

мы можем сделать важный вывод: при инвариантности распределения выборки по отношению к перестановкам упорядоченная статистика и ранговый вектор статистически независимы (теорема Гаека).

Прямым следствием этого факта является равенство соответствующих условных и безусловных распределений:

$$f(\{R_i\})=\frac{1}{N!}=f(\{R_i\}\,|\,\{x_{(R)}\}), \tag{5.2.6}$$

$$f(\{x_{(R)}\})=N!\;\; p(x_{(1)},\ldots,\; x_{(N)})=f(\{x_{(R)}\}\,|\,\{R_i\}). \tag{5.2.7}$$

Другим важным следствием статистической независимости упорядоченной выборки и рангового вектора является следующая теорема (Гаек и Шидак [1]).

При инвариантности распределения выборки к перестановкам для любой статистики $t(x_1, x_2, \ldots, x_N)$ выполняется соотношение

$$E\,[t(x_1,\ldots,\; x_N)\;\{R_i\}]=E\,[t(x_{(R_1)},\ldots,\; x_{(R_N)})]. \tag{5.2.8}$$

Действительно, так как $x_i=x_{(R_i)}$

$$E\,[t(x_1,\ldots,\; x_N)/\{R_i\}]=E\,[t(x_{(R_1)},\ldots,\; x_{(R_N)})/\{R_i\}],$$

и в силу независимости между $\{x_{(R)}\}$ и $\{R_i\}$ имеем (5.2.8).

Далее, для независимых выборочных значений с одинаковыми и симметричными относительно нуля распределениями $(p(x)=p(-x))$ можно ввести знаковую статистику и упорядочивание абсолютных значений. Определим знаковую функцию как

$$\operatorname{sign}(x)=\begin{cases}+1, & x>0,\\ 0, & x=0,\\ -1, & x<0,\end{cases} \tag{5.2.9}$$

и знаковую статистику $\operatorname{sign}(x_i)$. Тогда имеем следующую совокупность векторов, связанных с исходной выборкой:

знаковый вектор $\{\operatorname{sign} x_i\}$,

вектор абсолютных значений $\{|x_i|\}$,

упорядоченная статистика абсолютных значений $\{|x|_{(R_i^+)}\}$,

ранговый вектор абсолютных значений $\{R_i^+\}$.

Связь между компонентами этих векторов выражается очевидным равенством

$$x_i=\operatorname{sign} x_i\cdot|x|_{(R_i^+)}\,. \tag{5.2.10}$$

Из теоремы Гаека следует, что случайные векторы $\{\operatorname{sign} x_i\}$, $\{|x|_{(R_i)}\}$, $\{R_i^+\}$ тоже независимы. Их распределения равны соответственно:

$$P(\{\operatorname{sign} x_i\})=\left(\frac{1}{2}\right)^N, \tag{5.2.11}$$

$$P(\{R^+\}) = \frac{1}{N!}, \tag{5.2.12}$$

$$P(\{|x|_{(R^+)}\}) = 2^N \cdot N! \cdot \prod_{R=1}^{N} p(|x|_{(R)}). \tag{5.2.13}$$

Кроме того, очевидно, что при статистической независимости двух случайных величин X и Y их упорядоченные статистики и ранговые векторы, независимые для каждой из величин в теореме Гаека, независимы между собой.

§ 5.3. РАСПРЕДЕЛЕНИЕ РАНГОВОГО ВЕКТОРА ПРИ ОТСУТСТВИИ ИНВАРИАНТНОСТИ К ПЕРЕСТАНОВКАМ

Неинвариантность рангового вектора к перестановкам (т. е. неравновероятность различных перестановок чисел R_1, R_2, ..., R_N) может происходить либо непосредственно из-за несимметричности распределения выборки относительно перестановок аргументов, либо (при симметричности распределения) благодаря специально выбранному способу упорядочивания, обеспечивающего нарушение равновероятности различных реализаций рангового вектора при отходе от нулевой гипотезы.

Рассмотрим сначала случай, когда распределение $q(x_1, x_2, ..., x_N)$ выборки не обладает инвариантностью к перестановкам. Вероятность конкретной реализации $\{R_i\}$ рангового вектора, как и ранее, выражается интегралом

$$P_q(\{R_i\}) = \int_{S_R} q(x_1, ..., x_N)\, dx_1 ... dx_N, \tag{5.3.1}$$

где интегрирование ведется по области S_R тех значений $x_1, ..., x_N$, которые при упорядочивании дадут заданный вектор $\{R_i\}$. Однако из-за несимметричности q выражение (5.3.1) не может теперь быть приведено к виду (5.2.3).

Непосредственное вычисление многомерного интеграла (5.3.1) обычно затруднительно; поэтому представляет большой интерес приведение его к более удобной форме. Зададим произвольную плотность $p(x_1, ..., x_N)$, удовлетворяющую только двум ограничениям:

а) p является функцией, симметричной относительно перестановок аргументов; б) $p>0$ везде, где $q>0$ Помножив и разделив подынтегральное выражение (5.3.1) на $N!$ $p(x_1, ..., x_N)$, имеем:

$$P_q(\{R_i\}) = \frac{1}{N!} \int_{S_R} \left[\frac{q(x_1, ..., x_N)}{p(x_1, ..., x_N)}\right] \cdot N!\, p(x_1, ..., x_N)\, dx_1 ... dx_N. \tag{5.3.2}$$

Благодаря симметрии $p(x_1, \ldots, x_N) = p(x_{(1)}, \ldots, x_{(N)})$; далее, так как $N!\, p(x_{(1)}, \ldots, x_{(N)})$ есть распределение упорядоченной статистики, отличное от нуля только внутри области S_R, то (5.3.2) равно

$$P_q(\{R_i\}) = \frac{1}{N!} \int \left[\frac{q(x_1, \ldots, x_N)}{p(x_1, \ldots, x_N)} \right] \cdot f_p(\{x_{(R)}\} \mid \{R_i\})\, dx_{(1)} \ldots dx_{(N)}. \quad (5.3.3)$$

Теперь (5.3.3) выглядит как условное математическое ожидание статистики $t(x_1, \ldots, x^N) = q(x_1, \ldots, x^N)/p(x_1, \ldots, x_N)$, и в силу свойства (5.2.8) окончательно получаем:

$$P_q(\{R_i\}) = \frac{1}{N!} E_p \left[\frac{q(x_{(R_1)}, \ldots, x_{(R_N)})}{p(x_{(R_1)}, \ldots, x_{(R_N)})} \right]. \quad (5.3.4)$$

Таким образом, доказана теорема Хёфдинга [1]: вероятность заданной реализации рангового вектора при распределении q пропорциональна математическому ожиданию отношения правдоподобия q/p, усредняемого по множеству всех упорядоченных статистик выборки из распределения p, отвечающих заданному ранговому вектору.

Эта теорема может быть обобщена на случай совокупности ранговых векторов нескольких выборок или совокупности векторов многомерной выборки (Фрэйзер [1]):

$$P_q(\{R_{1i}\}, \{R_{2i}\}, \ldots, \{R_{ki}\}) =$$

$$= \frac{1}{N_1!\, N_2! \ldots N_k!} E_p \left[\frac{q(x_{1(R_1)}, \ldots, x_{1(R_{N_1})}; \ldots; x_{k(R_1)}, \ldots, x_{k(R_{N_k})})}{p(x_{1(R_1)}, \ldots, x_{1(R_{N_1})}; \ldots; x_{k(R_1)}, \ldots, x_{k(R_{N_k})})} \right]. \quad (5.3.5)$$

Здесь $\{R_{ji}\}$ ранговый вектор j-й выборки объема N_j при ее отдельном (!) упорядочивании

Перейдем теперь к обобщению теоремы Хёфдинга на случай, когда распределение исходной выборки симметрично относительно своих переменных, а неинвариантность распределения рангового вектора обязана своим происхождением соответствующему способу ранжировки. Начнем с двувыборочных задач, для которых подготовительными операциями являются: а) образование из выборок $\{x_{ji}\}$ ($j=1, 2$, $i=1, 2, \ldots, N_j$) смешанной выборки $\{w_l\}$ ($l=1, 2, \ldots, N_1+N_2$); б) упорядочивание смешанной выборки; в) отбор рангов, принадлежащих одной из исходных выборок. Пусть для простоты распределения q не только симметрично, но и факторизуемо (т. е. $\{x_{ji}\}$ — независимы). Тогда распределение рангового вектора $\{R_{ai}\}$ выразится как

$$P_q(\{R_{\alpha l}\}) = \frac{1}{\binom{N_1+N_2}{N_\alpha}} \cdot E_p\left\{\prod_{l=1}^{N_1+N_2}\left[\frac{q(w_{(R_l)})}{p(w_{(R_l)})}\right]^{Z_{\alpha l}}\right\}, \qquad (5.3.6)$$

где $Z_{\alpha l}$ — индикатор, введенный в § 5.1, а $w_{(R)}$ — R-я порядковая статистика выборки из распределения $p(w_\nu)$.

В ряде случаев информацию о ранговом векторе $\{R_{\alpha l}\}$ удобнее представить в виде вектора-индикатора $\{Z_{\alpha l}\} = (Z_{\alpha 1}, \ldots, Z_{\alpha(N_1+N_2)})$. Если плотности распределений $f(x)$ и $h(y)$ заданы, выборочные значения независимы, то

$$P(\{Z_{\alpha l}\}|f, h) = N_1!\, N_2! \int\ldots\int \prod_{l=1}^{N_1+N_2} f^{Z_{Xl}}(w_{(l)}) \cdot h^{1-Z_{Xl}}(w_{(l)})\, dw_{(l)} \cdot$$

$$-\infty \leqslant w_{(1)} \leqslant w_{(N_1+N_2)} \leqslant \infty. \qquad (5.3.7)$$

Переход к одновыборочным задачам (типа задач на симметрию распределения) несколько осложнен тем, что при образовании смешанной выборки из модулей выборочных значений (при проверке симметрии относительно нуля) число компонент каждого из ранговых векторов $\{R^+\}$ и $\{R^-\}$ является случайным (хотя суммарное число компонент обоих векторов равно объему выборки N). Однако для фиксированного числа M выборочных значений одного знака можно пользоваться формулами (5.3.6) и (5.3.7), заменив в них N_1+N_2 на N и N_α на M_α.

§ 5.4. О СТАТИСТИЧЕСКОЙ СВЯЗИ МЕЖДУ НАБЛЮДЕНИЕМ И ЕГО РАНГОМ

Статистическая связанность наблюдения x_i и его ранга R_i проявляется уже в том, что, встретившись со сравнительно большим выборочным значением, мы ожидаем, что оно получит сравнительно высокий ранг. Проведем, однако, более конкретное рассмотрение статистической связи случайных величин X и R.

Прежде всего, наличие такой связи выражается в нефакторизуемости совместного распределения X и R или в неравенстве соответствующих условных и безусловных распределений. Пусть, для конкретности, наблюдаемые величины $X_1, X_2, \ldots, X_N$ независимы и одинаково распределены с плотностью $p(x)$. Безусловное распределение величины X, (когда в качестве условия фигурирует знание ранга R_i), очевидно, характеризуется плотностью

$$f(x_i) = p(x_i). \qquad (5.4.1)$$

Если же ранг R_t известен, то выборочное значение x_t при этом условии является порядковой статистикой, т. е. $f(x_t/R)=f(x_{(R)})$, и, в соответствии с (3.2 8),

$$f(x_t\ R)=\frac{N!}{(R-1)!\,(N-R)!}\cdot F^{R-1}(x_t)\,[1-F(x_t)]^{N-R}\cdot p(x_t). \tag{5.4.2}$$

Сравнение (5.4.1) и (5.4.2) дает $f(x_t)\neq f(x_t/R)$, что и означает наличие зависимости между X и R. Получим остальные распределения, характеризующие ситуацию. Так как безусловное распределение ранга в наших условиях равномерно (см. § 5 2):

$$P(R)=\frac{1}{N}, \tag{5.4.3}$$

то совместное распределение $f(x_t, R)$ равно

$$f(x_t, R)=f(x_t/R)\cdot P(R)=$$

$$=\frac{(N-1)!}{(R-1)!\,(N-R)!}F^{R-1}(x_t)[1-F(x_t)]^{N-R}p(x_t). \tag{5.4.4}$$

(Отсюда, кстати, можно формально получить (5.4.1): $f(x_t)=\sum_R f(x_t,\ R)=p(x_t)$). И, наконец, условное распределение ранга R при известном x_t будет равно

$$P(R\ x_t)=f(x_t,\ R)/f(x_t)=$$

$$=\frac{(N-1)!}{(R-1)!\,(N-R)!}F^{R-1}(x_t)[1-F(x_t)]^{N-R}. \tag{5.4.5}$$

Как видим, (5.4.4) не факторизуется по переменным x_t и R и $P(R)\neq P(R/x_t)$. Итак, выборочное значение и его ранг являются статистически связанными случайными величинами. Характер связи между X и R полностью определяется приведенными выше распределениями, интересно, однако, рассмотреть некоторые частные характеристики статистической связи между этими случайными величинами (Ф. П. Тарасенко и В. П. Шуленин [1]).

Начнем с линий регрессии, $E(R/x_t)$ и $E(x_t/R)$.

$$E(R\ x_t)=\sum_R R\cdot P(R/x_t)=$$

$$=\sum_{R=1}^{N}\frac{R^2}{NF(x_t)}\cdot\binom{N}{R}F^R(x_t)\cdot[1-F(x_t)]^{N-R}=$$

$$=\frac{1}{NF(x_t)}\cdot\langle R^2\rangle_D,$$

где $\langle R^2\rangle_D$ — средний квадрат величины, распределенной биномиально, который можно вычислить, например, с помощью производящей функции моментов. Окончательно имеем:

$$E(R/x_i)=1+(N-1)F(x_i). \qquad (5.4.6)$$

Вычислим теперь $E(x_i/R)$.

$$E(x_i/R)=\int x_i\cdot f(x_i/R)\,dx_i=$$

$$=\frac{N!}{(R-1)!\,(N-R)!}\int x_i F^{R-1}(x_i)\,[1-F(x_i)]^{N-R}\,p(x_i)\,dx_i=$$

$$=\frac{N!}{(R-1)!\,(N-R)!}\int_0^1 F^{-1}(y)\cdot y^{R-1}\cdot(1-y)^{N-R}\,dy. \qquad (5.4.7)$$

Дальнейшие аналитические выкладки возможны лишь при конкретизации $F(x)$. Пусть, например, $F(x)=x$, $x\in[0,\ 1]$. Тогда

$$E(x_i\ R)=\frac{N!}{(R-1)!\,(N-R)!}\int_0^1 y^R(1-y)^{N-R}dy=\frac{R}{N+1}. \qquad (5.4.8)$$

На рис. 5.4.1 приведены линии регрессии (5.4.6) и (5.4.8) для равномерного распределения. Наличие связи между x_i и R выражается в том, что линии регрессии неперпендикулярны. Статистичность, неоднозначность этой связи приводит к несовпадению линий регрессии (известно, что линии регрессии сливаются, если связь носит функциональный характер). Однако бросается в глаза тот факт, что при увеличении объема выборки угол между линиями регрессии неограниченно убывает, что говорит об усилении связи при возрастании N.

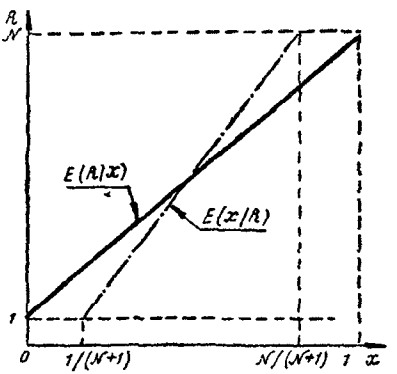

Рис. 5.4.1

Рассмотрим другую меру статистической связи — коэффициент корреляции ρ_{XR} между x_i и R. По определению,

$$\rho_{XR}(N)=\frac{E\,[(X-EX)(R-ER)]}{[DX]^{1/2}\,[DR]^{1/2}}. \qquad (5.4.9)$$

Учитывая, что $ER=\dfrac{N+1}{2}$; $DR=\dfrac{N^2-1}{12}$, имеем:

$$\rho_{XR}(N)=\frac{E(XR)-\frac{N+1}{2}E(X)}{[DX]^{1/2}\cdot\sqrt{\frac{N^2-1}{12}}}. \qquad (5.4.10)$$

Используя (5.4.6), для $E(XR)$ легко получить:

$$E(XR)=\int x\left\{\sum_R R\cdot\frac{(N-1)!}{(R-1)!\,(N-R)!}F^{R-1}(x)[1-F(x)]^{N-R}\right\}dF(x)=$$
$$=\int x\,\{1+(N-1)\,F(x)\}\,dF(x). \qquad (5.4.11)$$

Подставив (5.4.11) в (5.4.10) и приводя подобные члены, окончательно имеем:

$$\rho_{XR}(N)=\sqrt{\frac{N-1}{N+1}}\left[\sqrt{\frac{12}{DX}}\cdot\int x\left[F(x)-\frac{1}{2}\right]dF(x)\right]=$$
$$=\sqrt{\frac{N-1}{N+1}}\cdot\rho(F). \qquad (5.4.12)$$

Соотношение (5.4.12) интересно тем, что коэффициент корреляции факторизовался: один множитель зависит только от объема выборки (и стремится к единице при $N\to\infty$), второй множитель, $\rho(F)$ определяется только распределением $F(x)$ исходной выборки. Еще более интересным оказывается то, что $\rho(F)$ зависит только от в и д а распределения, но не от его масштаба или сдвига*; т. е.

$$\rho(F(x))=\rho\left(F\left(\frac{x-a}{\sigma}\right)\right) \qquad (5.4.13)$$

при любых a и σ.

В самом деле

$$\rho\left(F\left(\frac{x-a}{\sigma}\right)\right)=\frac{\sqrt{12}}{\sigma}\int x\left[F\left(\frac{x-a}{\sigma}\right)-\frac{1}{2}\right]dF\left(\frac{x-a}{\sigma}\right)=$$
$$=\frac{\sqrt{12}}{\sigma}\int(\sigma t+a)\left[F(t)-\frac{1}{2}\right]dF(t)=$$
$$=\frac{\sqrt{12}}{\sigma}\left\{\sigma\int x\left[F(t)-\frac{1}{2}\right]dF(t)+a\int\left[F(t)-\frac{1}{2}\right]dF(t)\right\}=$$
$$=\rho(F(x))+\frac{a\sqrt{12}}{\sigma}\int\left[F(t)-\frac{1}{2}\right]dF(t)=\rho(F(x)), \qquad (5.4.14)$$

* Разумеется, имеются в виду лишь распределения, для которых существуют интегралы, входящие в определение коэффициента корреляции.

так как последний интеграл в (5.4.14) равен нулю. Вычисления $\rho(F)$ для некоторых типов распределений дают следующие результаты (Стьюарт [1], Кендалл [1])

для равномерного распределения

$$\rho_R(F)=1, \tag{5.4.15}$$

для нормального распределения

$$\rho_N(F)=\sqrt{\frac{3}{\pi}}=0{,}9772, \tag{5.4.16}$$

для семейства гамма-распределений, элемент вероятности которых записывается в виде

$$dF(x)=\frac{1}{\Gamma(m)}\cdot e^{-x}x^{m-1}dx, \tag{5.4.17}$$

$$\rho_\Gamma(F)=\sqrt{\frac{3m}{\pi}}\cdot\frac{\Gamma\left(m+\frac{1}{2}\right)}{\Gamma(m+1)}=\frac{\sqrt{3m}}{2^m}\cdot\frac{(2m-1)!!}{m!}. \tag{5.4.18}$$

В связи с тем, что вычисления линий регрессии доводятся до конца лишь в частном случае равномерного распределения выборочных значений, а коэффициент корреляции вскрывает лишь линейную компоненту статистической зависимости и является исчерпывающим показателем связи лишь для компонент нормального двумерного распределения, возникает некоторое сомнение в общности полученных выше результатов. Рассмотрим поэтому более общую количественную меру связи двух случайных величин — количество информации в одной из них о другой — при произвольном распределении $p(x)$. Подставив в функционал количества информации

$$I(X,R)=\sum_{R=1}^{N}\int_X f(x,R)\ln\frac{f(x,R)}{f(x)P(R)}dx. \tag{5.4.19}$$

распределения (5.4.1), (5.4.3) и (5.4.4), производя замену $F(x)=t$ и расписав логарифм произведения как сумму логарифмов, имеем:

$$\begin{aligned}I(X,R)=&\sum_{R}^{N}\ln\frac{N!}{(R-1)!\,(N-R)!}\cdot C_{NR}\int_0^1 t^{R-1}(1-t)^{N-R}dt+\\&+\sum_{R}^{N}C_{NR}\int_0^1 t^{R-1}(1-t)^{N-R}\ln t^{R-1}dt+\\&+\sum C_{NR}\int_0^1 t^{R-1}(1-t)^{N-R}\ln(1-t)^{N-R}dt,\end{aligned} \tag{5.4.20}$$

где $C_{NR}=(N-1)!/[(R-1)!(N-R)!]$. Первый интеграл равен $1/(NC_{NR})$ (см. (3.3.2)); второй и третий выражаются через бета- и пси-функции (см. Рыжик и Градштейн [1], (4.253.1)), и после несложных преобразований получаем:

$$I(X,R)=\frac{1}{N}\sum_{R}^{N}\ln\frac{N!}{(R-1)!\,(N-R)!}-\frac{1}{N}\sum_{R}^{N}\Big[(N-1)\sum_{k=1}^{V}\frac{1}{k}-$$
$$-(R-1)\sum_{k=1}^{R-1}\frac{1}{k}-(N-R)\sum_{k=1}^{N-R}\frac{1}{k}\Big]. \qquad (5.4.21)$$

С помощью перегруппировки слагаемых при двойном суммировании можно показать, что

$$\frac{1}{N}\sum_{R=1}^{N}(N-1)\sum_{k=1}^{N}\frac{1}{k}=(N-1)\sum_{k=1}^{N}\frac{1}{k};$$

$$\frac{1}{N}\sum_{R=1}^{N}(R-1)\sum_{k=1}^{R-1}\frac{1}{k}=\frac{1}{N}\sum_{R=1}^{N}(N-R)\sum_{k=1}^{N-R}\frac{1}{k}=$$
$$=\sum_{s=1}^{N-1}\frac{1}{S}\sum_{k=s}^{N-1}k=\frac{N-1}{2}\sum_{k=1}^{N-1}\frac{1}{k}-\frac{1}{4}\frac{(N-1)(N-2)}{N};$$

$$\frac{1}{N}\sum_{R=1}^{N}\ln(N-1)!=\sum_{k=1}^{N-1}\ln k;$$

$$\frac{1}{N}\sum_{R=1}^{N}\ln(R-1)!=\frac{1}{N}\sum_{R=1}^{N}\ln(N-R)!=\frac{1}{N}\sum_{k=1}^{N-1}k\ln(N-k)=$$
$$=\frac{1}{N}\sum_{k=1}^{N-1}(N-k)\ln k.$$

Окончательно получаем:

$$I(X,R)=\ln N-\left(\sum_{k}^{N-1}\ln k+\frac{N-1}{2}-\frac{2}{N}\sum_{k}^{N-1}k\ln k\right). \qquad (5.4.22)$$

Дальнейшее упрощение (5.4.22) представляется невозможным, поэтому рассмотрим асимптотику полученных рядов. С помощью формулы суммирования Эйлера-Маклорена (см. де Брёйн [1]) можно получить, что при $N\to\infty$

$$\sum_{k=1}^{N}k\ln k=\frac{N^2}{2}\ln N-\frac{N^2}{4}+\frac{N}{2}\ln N=\frac{1}{4}\ln N+0(1) \qquad (5.4.23)$$

и

$$\sum_{k=1}^{V}\ln k=N\ln(N+1)+\frac{1}{2}\ln(N+1)-(N+1)+$$

$$+\frac{1}{2}\ln 2\pi+0\left(\frac{1}{N}\right). \qquad (5.4.24)$$

Используя эти формулы, окончательно (с точностью до членов порядка $0(\ln N/N)$) имеем:

$$I(X,\ R)=\frac{1}{2}\ln N+\frac{1}{2}\ln\frac{e}{2\pi}. \qquad (5.4.25)$$

О качестве полученного приближения можно судить по таблице, в которой приведены точные (правый столбец) и вычисленные по формуле (5.4.25) значения $I(X,\ R)$ для различных N:

2	0,193	0,065	200	2,236	2,232
5	0,522	0,388	500	2,691	2,691
10	0,812	0,735	1000	3,036	3,038
20	1,124	1,082	2000	3,382	3,384
50	1,558	1,539	5000	3,840	3,842
100	1,895	1,896	10000	4,186	4,189

Из (5.4.25) следует, что количество информации в R о X неограниченно возрастает с объемом выборки и что эта зависимость не только не связана с параметрами сдвига и масштаба распределения, но и одинакова для всех непрерывных распределений.

Наличие и усиление с ростом N статистической связи между выборочным значением и его рангом является основой для заключения о том, что в принципе возможно построение статистических процедур, использующих только ранги и по своим качествам асимптотически приближающихся к процедурам на выборочных значениях.

Однако эта статистическая связь носит характер необходимого, но не достаточного условия для такого построения. В самом деле, переход к рангам (при непосредственном упорядочивании выборки) исключает зависимость от сдвига и масштаба, и, если именно они являются информативными параметрами задачи, возможность ее решения ранговой процедурой отпадает. Именно поэтому при желании использовать ранговую процедуру для решения некоторой задачи необходимо прежде всего найти такой способ упорядочивания, при котором информативный параметр задачи (т. е. параметр различия между конкурирующими гипотезами) влиял бы на распределение рангов. Собственно, такая ситуация объясняется тем, что из наличия связей между величинами Θ и X и X и R еще не следует наличие связи между Θ и R. Лишь при специальном способе упорядочивания, т. е. при специальной организации последовательности пар связанных величин

$(\Theta, X) \to (X, Z) \to (Z, R)$ первая и последняя величины оказываются связанными в результате принятых мер.

Если такие меры приняты, то остается вопрос, как именно использовать ранги, чтобы получить статистику, асимптотически эквивалентную заданной статистике $S(x_1, \ldots, x_N)$. Ответ на этот вопрос дает формула (3.5.9), которая позволяет утверждать, что искомая статистика имеет вид $S(F^{-1}(\frac{1}{N+1}), \ldots, F^{-1}(\frac{N}{N+1}))$. (Отметим, что получение такой ранговой статистики возможно лишь при знании функции $F(x)$).

§ 5.5. СПОСОБЫ ОБРАЩЕНИЯ СО СВЯЗКАМИ

В отличие от выборок непрерывных случайных величин, выборки дискретных величин могут содержать связки, т. е. совокупности совпадающих значений. В таком случае при упорядочивании возникает вопрос: как распределить ранги среди членов связки? Поскольку ряд статистических процедур использует только ранги, то от того, как именно обрабатываются связки, могут существенно зависеть свойства этих процедур.

Непосредственное применение формулы (5.1.2) для вычисления ранга приводит к тому, что при наличии связки всем ее членам приписывается одинаковый и наибольший возможный ранг. Это может оказаться неудобным; и уж если примириться с одинаковостью рангов внутри связки, то из общих соображений более естественно потребовать, чтобы при сохранении разницы между рангами ближайших соседей связки, равной числу ее членов, самим членам связки присваивался ранг, средний между минимальным и максимальным возможными. В связи с этим вместо (5.1.1) введем новую функцию сравнения

$$C'(t) = \begin{cases} 1, & t>0, \\ \frac{1}{2}, & t=0, \\ 0, & t<0, \end{cases} \tag{5.5.1}$$

с помощью которой ранги для значений, принадлежащих выборке дискретной величины, будем вычислять по формуле:

$$R_i' = \frac{1}{2} + \sum_{k=1}^{N} C'(x_i - x_k). \tag{5.5.2}$$

Для одиночных значений X (5.5.2) дает тот же результат, что и (5.1.2), а при наличии связок их членам присваивается

общий, средний ранг. Для отличия рангов, вычисляемых по формуле (5.5.2), от обычных введем название «мидрангов» (по аналогии с английским термином midrank). Мидранг можно еще интерпретировать как «среднее значение ранга для членов связки, если бы они оказались различимыми».

На практике часто удобнее работать не с мидрангами (5.5.2), а с их линейной функцией

$$W_i=2\left[R'_i-\frac{N+1}{2}\right]=\sum_{k=1}^{N}\operatorname{sign}(x_i-x_k), \qquad (5.5.3)$$

где

$$\operatorname{sign}(t)=\begin{cases}1, & t>0,\\ 0, & t=0,\\ -1, & t<0.\end{cases} \qquad (5.5.4)$$

Достоинством W-рангов (5.5.3) является то, что они не принимают дробных значений внутри связок и что их сумма равна нулю.

До сих пор мы рассматривали регулярные способы обработки связок, при которых всем членам связки присваивается один и тот же ранг. Можно ли, сохранив равноправность членов внутри связки, присвоить им все же различные ранги? Такую возможность предоставляет хорошо известный даже из житейской практики способ «распределения по жребию». Говоря точнее, пусть $x_{i1}=x_{i2}=x_{i3}=\ldots=x_{ir}$ — члены некоторой связки r одинаковых значений x. Произведем некоторый случайный эксперимент с $r!$ возможными и равновероятными исходами. Установим взаимно однозначное соответствие между исходами этого эксперимента и всевозможными упорядочиваниями r объектов. Будем присваивать ранги членам связки в том порядке, который предписывается исходом эксперимента. При этом равноправность членов связки соблюдается в форме равновероятности всевозможных упорядочиваний.

Следует особо подчеркнуть, что случайное присвоение рангов внутри связки сохраняет статистические свойства рангового вектора при инвариантности распределения к перестановкам в одновыборочных значениях и при одинаковости распределений в многовыборочных. В самом деле, в указанных случаях реализации ранговых векторов оказываются попрежнему равновероятными, как это было для непрерывного случая (см. § 5.2). Поэтому для этих гипотез статистические ранговые процедуры, синтезированные для непрерывных распределений, можно применять к дискретным выборкам — при условии случайного упорядочивания внутри связок. Это замечание приобретает особую важность в связи с тем, что

при практических измерениях непрерывных величин всегда приходится иметь дело с дискретными значениями — просто в силу ограниченной точности приборов.

Для случая двувыборочных задач Леман [2] доказал следующую теорему. Пусть $x_1, x_2, \ldots, x_{N_1}; x_{N_1+1}, x_{N_1+2}, \ldots, x_{N_1+N_2}$ — независимы; пусть все x_i $(i=1, 2, \ldots, N_1)$ имеют одинаковые функции распределения $F(x)$, а все x_{N_1+j} $(j=1, 2, \ldots, N_2)$ — функции $G(x)$. Если $A=A(\{R_i\})$ — случайное событие, осуществление которого зависит только от рангового вектора α-х компонент смешанной статистики, и если ранги в связках присваиваются случайным образом, то для любых разрывных $F(x)$ и $G(x)$ существуют соответствующие непрерывные F^* и G^*, такие, что

$$P_{F^*G^*}(A)=P_{FG}(A), \tag{5.5.5}$$

и $F^*=G^*$, если и только если $F=G$.

Приведем один из способов построения таких функций F^* и G^*, которые обеспечивают соблюдение равенства (5.5.5). Пусть F и/или G имеют разрывы на счетном множестве точек $a_1, a_2, \ldots$ Построим сначала пару функций F_1, G_1 следующим путем. Разрежем графики F и G в точке a_1 и раздвинем их части на расстояние $\delta/2$ друг от друга. Внутри образовавшегося просвета проведем прямые, соединяющие концы раздвинутых графиков; получившиеся в результате этих манипуляций функции обозначим через F_1 и G_1, соответственно. Следующий такой разрез проведем в точке, соответствующей теперь разрыву, ранее находящемуся в точке a_2. Раздвинем полученные половинки на расстояние $\delta/2^2$; внутри нового интервала снова проведем прямые; получившиеся функции обозначим через F_2 и G_2. Циклически повторяя операции «разрез по a_k — раздвижение на $\delta/2^k$ — соединение прямыми», мы получим последовательность пар функций (F_1, G_1), (F_2, G_2), $\ldots$, (F_k, G_k), $\ldots$ Эта последовательность пар функций имеет три интересных особенности. Во-первых, подмена k-го разрыва любой из функций F или G равномерным в интервале $\delta/2^k$ распределением эквивалентно тому, что ранги в k-й связке приписываются случайным образом с одинаковой вероятностью. Во-вторых, при указанном способе построения F_k и G_k распределение рангового вектора остается неизменным на каждом шаге. В-третьих, последовательности F_k и G_k даже при бесконечном числе разрывов сходятся к некоторым непрерывным F^* и G^* соответственно в силу сходимости суммы $\Sigma 2^{-k}$. Тем самым мы доказали первую часть теоремы до равенства (5.5.5) включительно. Заключительное утвержде-

ние является очевидным для описанной выше процедуры построения F^* и G^*.

Легко видеть, что обобщение теоремы Лемана на случай одновыборочных и многовыборочных задач можно провести непосредственно.

Известны и другие способы обращения со связками. Перечислим основные из этих методов (учитывая, что способы: а) использование мидрангов и б) рандомизация рангов внутри связки — уже были изложены).

в) Взять из каждой связки только по одному наблюдению и соответственно уменьшить объем выборки. (Заметим, что потери в мощности при этом в ряде случаев могут быть меньшими, чем при методах б) и г) Хемельрик [1], Паттер [1]).

г) Приписать половине членов связки наименьший возможный ранг, другой половине — наибольший. (Разновидность этого способа предписывает, например, при знаковом тесте половину нулей в связке считать положительными, а другую половину — отрицательными.

д) Оценить границы вероятности ошибки первого рода. Идея этого метода состоит в том, чтобы сначала обработать связки таким способом, который должен дать наибольшую вероятность отвержения нулевой гипотезы, а затем так, чтобы отвержение было наименее вероятным (выбор конкретных способов обработки связок определяется прежде всего типом тестовой статистики). Затем следует оценить вероятности $\alpha_{\text{М}}$ и $\alpha^{\text{Б}}$ ошибок первого рода при этих способах, предполагая, что верна нулевая гипотеза. При любой другой обработке связок вероятность ошибок $\alpha(T)$ находится, очевидно, между ними: $\alpha_{\text{М}} \leqslant \alpha(T) \leqslant \alpha_{\text{Б}}$. Далее $\alpha_{\text{М}}$ и $\alpha_{\text{Б}}$ сравниваются с требуемым уровнем значимости α_0. Если обе они окажутся больше (или меньше) α_0, гипотеза H_0 отвергается (принимается); при $\alpha_{\text{М}} < \alpha_0 < \alpha_{\text{Б}}$ данный способ не дает решения задачи. Если условия эксперимента позволяют, следует увеличивать объем выборки, пока решение не будет достигнуто.

Каждый из указанных выше способов обращения со связками обладает определенными достоинствами и недостатками; однако последний способ представляется наиболее приемлемым с точки зрения полного использования всех наличных наблюдений.

§ 5.6. О СВОЙСТВАХ РАНГОВ КОМПОНЕНТ ДВУМЕРНОЙ ВЫБОРКИ

В § 5.3 было показано наличие статистической связи между выборочным значением x_i и его рангом R_i. Один из способов конкретного использования этой связи состоит в том, чтобы в определенных процедурах вместо самих выборочных значений использовать их ранги. Известным примером такой замены является использование коэффициента корреляции между рангами R и K двух величин X и Y вместо коэффициента корреляции между самими X и Y (так называемый ранговый коэффициент Спирмэна). Правда, иногда (см., например, Нетер [1]) высказывается мнение, будто аналогия между этими коэффициентами чисто внешняя; однако, учитывая связь между X и R, Y и K (§ 5.4), такое мнение можно оспаривать.

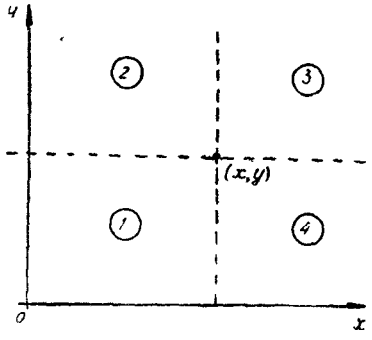

Рис. 5.6.1

Интересно было бы подвергнуть более детальному рассмотрению вопрос о связях между многомерными выборочными значениями и соответствующими многомерными совокупностями рангов. Наиболее полно эти связи могут быть описаны с помощью соответствующих функций распределения. Однако получение этих функций наталкивается на ряд затруднений. Например, Кумару [1] удалось даже в двумерном случае выразить эти распределения лишь в виде рекуррентных соотношений.

Тем не менее, можно предложить вывод искомых распределений для двумерного случая, дающий результат не в рекуррентной форме, а в явном виде. Выберем на плоскости XOY некоторую точку (x, y) (см. рис. 5.6.1). Прямые, параллельные осям координат, рассекают плоскость XOY на четыре квадранта, пронумерованные на рисунке по часовой стрелке. Вероятность P_k того, что некоторая выборочная пара значений (x_i, y_i) окажется в k-м квадранте, равна соответственно

$$\begin{aligned} p_1 &= F(x, y), \\ p_2 &= F(x) - F(x, y), \\ p_4 &= F(y) - F(x, y), \\ p_3 &= 1 - F(x) - F(y) + F(x, y). \end{aligned} \tag{5.6.1}$$

Пусть теперь разбиение на квадранты произведено через выборочную точку $(x_{(R)}, x_{(K)})$, которой соответствует пара R и K рангов, присвоенных данным выборочным значениям при упрядочивании по каждой координате в отдельности. При этом между числами m_1, m_2, m_3 и m_4 выборочных точек в каждом из квадрантов выполняются соотношения

$$\begin{aligned} m_1+m_2+m_3+m_4&=N-1,\\ m_1+m_2\qquad\quad&=R-1,\\ m_1\qquad\quad+m_4&=K-1. \end{aligned} \tag{5.6.2}$$

Для заданного размещения выборочных точек по квадрантам вероятность осуществления равна

$$\frac{(N-1)!}{m_1!\,m_2!\,m_3!\,m_4!}\,p_1^{m_1}p_2^{m_2}p_3^{m_3}p_4^{m_4},$$

а так как в общем случае при фиксированных R и K система (5.6.2) оставляет возможность некоторого изменения чисел m_1, m_2, m_3 и m_4, то условную вероятность осуществления пары (R, K) при заданных (x_i, y_i) можно записать как

$$P[R, K)|(x_i, y_i)] = \sum_{M}\frac{(N-1)!}{m_1!\,m_2!\,m_3!\,m_4!}\,p_1^{m_1}p_2^{m_2}p_3^{m_3}p_4^{m_4}. \tag{5.6.3}$$

Остается указать конкретно множество M, по которому следует вести суммирование. Это легко сделать, учитывая, что три неизвестных в системе (5.6.2) могут быть выражены через четвертое (например, m_1) и что все числа m_k неотрицательны. Имея, следовательно, соотношения

$$\begin{aligned} m_2&=R-1-m_1\geqslant 0,\\ m_4&=K-1-m_1\geqslant 0,\\ m_3&=(N+1)-(K+R)+m_1\geqslant 0, \end{aligned} \tag{5.6.4}$$

получаем, что суммирование ведется по таким $m_1\in$M, что

$$\begin{aligned} \min(N-K,\ N-R)\geqslant m_1\geqslant[(K+R)-(N+1)]\cdot\\ \cdot C[(K+R)-(N+1)], \end{aligned} \tag{5.6.5}$$

где $C(t)=1$ при $t\geqslant 1$ и $C(t)=0$ при $t<0$.

Окончательно имеем

$$\begin{aligned} P[(R, K)|(x_i, y_i)] =\\ =\sum_{m_1\in M}\frac{(N-1)!}{m_1!\,(R-1-m_1)!\,[(N+1)-(K+R)+m_1]!\,(K-1-m_1)!}\times\\ \times p_1^{m_1}\cdot p_2^{R-1-m_1}\cdot p_3^{(N+1)-(K+R)+m_1}\cdot p_4^{K-1-m_1}. \end{aligned} \tag{5.6.6}$$

Учитывая, что безусловные распределения $f(x_{(R)}, y_{(K)})$ и $P(R, K)$ известны и равны соответственно

$$f(x_{(R)}, y_{(K)}) = p(x_{(R)}, y_{(K)}), \qquad (5.6.7)$$

$$P(R, K) = \frac{1}{N^2},$$

легко получить совместное распределение $f[(R, K), (x_{(R)}, y_{(K)})]$ и условное $P[(x_{(R)}, y_{(K)})/(R, K)]$, имея, таким образом, полное описание статистической связи между двумерными выборочными значениями и парами их частных рангов.

ГЛАВА VI

ЗАКОНЫ БОЛЬШИХ ЧИСЕЛ И ПРЕДЕЛЬНЫЕ ТЕОРЕМЫ КАК НЕПАРАМЕТРИЧЕСКИЕ ФАКТЫ

§ 6.1. ВВЕДЕНИЕ

Как уже неоднократно подчеркивалось, любое свойство выборки или функций от выборки (статистик), не зависящее от вида распределения, которому выборка подчинена, может быть использовано для построения непараметрических процедур. В главах III—V наше внимание было сосредоточено на поисках непараметрических свойств выборки. Однако исторически гораздо раньше, по существу, с самого начала развития теории вероятностей и статистики, были открыты непараметрические свойства некоторых статистик, таких, как относительные частоты и суммы выборочных значений. Правда, эти непараметрические свойства носят асимптотический характер, т. е. проявляются лишь при $N \to \infty$; однако практиков вполне удовлетворяет то, что при конечных N можно действовать так, как будто асимптотическое свойство уже имеет место, а погрешности или вероятности ошибок, связанные с таким подходом, можно оценить количественно. Очевидно, качество полученных таким образом процедур будет тем выше, чем больше N; понятно поэтому, что для достижения заданного показателя качества необходим «достаточно большой» объем выборки.

Асимптотические свойства статистик распадаются на два класса: 1) сходимость статистик к некоторым постоянным пределам; 2) сходимость распределений статистик к некоторым известным распределениям. Свойства первого класса получили название законов больших чисел, свойства второго класса — предельных теорем. Непараметрич-

ность этих свойств состоит в том, что они проявляются независимо от распределения исходной выборки; однако следует подчеркнуть, что наличие этих свойств доказано: а) лишь для вполне определенных типов статистик и б) лишь при выполнении определенных требований к свойствам наблюдаемой случайной величины. Многие выдающиеся математики приложили немало усилий к тому, чтобы ослабить эти требования; но практикам нелишне помнить, что иногда встречаются случаи, когда даже эти достаточно слабые ограничения не выполняются.

Законы больших чисел и предельные теоремы являются классическими разделами современной теории вероятностей. Поэтому в данной главе будут приведены лишь основные результаты; их доказательства читатель может найти, например, у Б. В. Гнеденко [3], А. Реньи [2] и Д. Фрейзера [1].

Известно, что доказательства законов больших чисел базируются на неравенствах Маркова и Чебышева. Желая подчеркнуть, что эти неравенства сами имеют непараметрический характер (т. е. являются справедливыми для непараметрического семейства функций распределения), мы обсудим их в отдельном параграфе.

§ 6.2. НЕРАВЕНСТВА МАРКОВА И ЧЕБЫШЕВА

Для непараметрической статистики весьма важен ответ на следующий вопрос: что можно сказать о поведении выборочных значений независимо от вида их распределения? Ответы типа «выборочные значения будут группироваться в основном в области наиболее вероятных значений» или «разброс выборочных значений связан с масштабным параметром распределения» верны, но неконкретны. Конструктивное значение имели бы лишь суждения, содержащие количественные характеристики. Поэтому большой интерес представляют знаменитые неравенства Маркова и Чебышева, являющиеся именно такими количественными утверждениями.

Теорема 6.2.1. (неравенство Маркова). Пусть X — положительная случайная величина с конечным средним $M=E(X)$. Тогда для любого $\mu>1$ справедливо неравенство

$$P(X\geqslant\mu M)\leqslant\frac{1}{\mu}. \tag{6.2.1}$$

Действительно, из $M=F(X)=\int_0^\infty xdF(x)$ следует, что

$$M \geqslant \int_{\mu M}^{\infty} x dF(x) \geqslant \mu M \int_{\mu M}^{\infty} dF(x) = \mu M (1 - F(\mu M)),$$

что и доказывает (6.2.1).

Теорема 6.2.2 (неравенство Чебышева). Пусть X — случайная величина с произвольным распределением, имеющим конечные среднее $M = E(X)$ и дисперсию $D(X) = \sigma^2$. Тогда для любого $\lambda > 1$

$$P(|X - M| > \lambda\sigma) \leqslant \frac{1}{\lambda^2}. \qquad (6.2.2)$$

Действительно, применив неравенство (6.2.1) к случайной величине $Y = (X - M)^2$ и положив $\mu = \lambda^2$, мы получим (6.2.2).

Пользуясь неравенствами Маркова и Чебышева, мы можем сделать важные заключения о поведении случайной величины, которые справедливы для широкого класса распределений. Например, если функция, обратная $F(x)$, однозначна, то (6.2.1) можно записать в виде

$$F^{-1}\left(1 - \frac{1}{\mu}\right) \leqslant \mu M,$$

откуда, в частности, следует, что верхний квартиль положительной случайной величины не может превышать ее учетверенное среднее, а медиана — удвоенное среднее. С другой стороны, неравенство Чебышева выражает замечательный непараметрический факт: для любой случайной величины можно указать конкретные пределы, за которые ее реализовавшееся значение не выйдет с известной вероятностью.

За получение столь общих и тем не менее конкретных суждений пришлось, однако, уплатить ценой некоторых ограничений на классы возможных распределений: для неравенства Маркова это — положительность X и существование $E(X)$, для неравенства Чебышева — существование первых двух моментов. Кроме того, полученные неравенства являются обычно весьма пессимистическими: для многих реальных распределений границы Чебышева намного превышают истинные границы, определяемые заданной вероятностью (см., например, А. Реньи [2], Ф. П. Тарасенко [3]).

Поэтому предпринимались попытки усилить неравенство Чебышева. Такую возможность предоставляет применение неравенства Маркова к различным положительным функциям разности $(X - M)$, (а не только к $(X - M)^2$, как при получении неравенства Чебышева).

Положим, например $Y = |X - M|^{\alpha}$, $\alpha > 0$, и обозначим $M_{\alpha} = E(|X - M|^{\alpha})$. Тогда из (6.2.1) имеем:

$$P(|X-M|>\lambda\sigma)\leqslant\frac{M_\alpha}{(\lambda\sigma)^\alpha}. \qquad (6.2.3)$$

Выбор α, минимизирующего правую часть (6.2.3), даст неравенство, усиленное по сравнению с неравенством Чебышева.

Аналогичные неравенства можно получить путем рассмотрения функции $Y=\exp[\varepsilon(X-M)]$. Введем обозначение

$$E(\exp[\varepsilon(X-M)])=B(\varepsilon). \qquad (6.2.4)$$

Применив (6.2.1), получим

$$P(\exp[\varepsilon(X-M)]>B(\varepsilon)e^t)<e^{-t}. \qquad (6.2.5)$$

Из монотонности экспоненциальной функции следует, что при $t>0$ и $\varepsilon>0$

$$P\left(X>M+\frac{t+\ln B(\varepsilon)}{\varepsilon}\right)<e^{-t}. \qquad (6.2.6)$$

Используя (6.2.6) и дополнительное условие ограниченности величины X, $|X|<K$, С. Н. Бернштейн [1] получил весьма сильное неравенство

$$P(|X-M|\geqslant\mu\sigma)\leqslant 2\exp\left[-\frac{\mu^2}{2\left(1+\frac{\mu K}{2\sigma}\right)^2}\right]. \qquad (6.2.7)$$

§ 6.3. ЗАКОНЫ БОЛЬШИХ ЧИСЕЛ

Во многих случаях статистики S_N, с помощью которых выносятся статистические решения, имеют аддитивную структуру, т. е. являются суммами некоторых случайных величин: $S_N=\sum_{i=1}^{N} x_i$. Оказывается, что при определенных условиях поведение таких статистик при неограниченном возрастании объема выборки N не зависит от вида распределения исходной выборки.

Совокупность таких утверждений об асимптотической сходимости статистик данного класса к определенным константам известна под названием законов больших чисел. Непараметричность законов больших чисел представляет интерес для непараметрической статистики, так как она непосредственно может использоваться и для синтеза и для анализа непараметрических процедур. Напомнив определение сходимости по вероятности (см. § 2.2), приведем основные формулировки законов больших чисел.

Теорема 6.3.1. (Теорема Бернулли). В последовательности независимых испытаний относительная частота события A сходится по вероятности к вероятности $P(A)$ события A при неограниченном увеличении числа испытаний.

Относительная частота m/N может быть представлена в форме среднего арифметического случайных величин x_i ($i=1, 2, ..., N$), принимающих значения 1 или 0 в зависимости от того, реализовалось или нет событие A в i-м испытании: $\frac{m}{N}=\frac{1}{N}\Sigma x_i=\frac{S_N}{N}$. Таким образом, теорема Бернулли утверждает, что в данном случае эмпирическое среднее сходится по вероятности к математическому ожиданию. Теорема Чебышева обобщает это утверждение на более общий случай произвольных независимых случайных величин:

Теорема 6.3.2. (Теорема Чебышева). Если $X_1, X_2, ..., X_N, ...$ — последовательность независимых случайных величин, имеющих конечные средние и дисперсии, то

$$\frac{S_N}{N} \xrightarrow{p} \frac{1}{N}\sum_{i=1}^{N} E(X_i).$$

Для одинаково распределенных величин можно доказать более сильный результат, не требующий существования дисперсий слагаемых.

Теорема 6.3.3. (Теорема Хинчина). Если случайные величины $X_1, X_2, ...$ независимы, одинаково распределены и имеют конечные средние M, то

$$\frac{S_N}{N} \xrightarrow{p} M.$$

Другое обобщение теоремы Чебышева дает

Теорема 6.3.4. (Теорема Маркова). Если последовательность случайных величин $X_1, X_2, ...$ такова, что при $N\to\infty$

$$\lim_{N\to\infty}\frac{1}{N^2}\cdot D(S_N)=0,$$

то

$$\frac{S_N}{N} \xrightarrow{p} \frac{1}{N}\sum_{i=1}^{N} E(X_i).$$

Важное значение имеют исследования Маркова, Бернштейна, Гнеденко, направленные на выяснение того, при каких условиях закон больших чисел имеет место для сумм зависимых случайных величин.

Теорема 6.3.5. (Теорема Гнеденко). Для того, чтобы для последовательности $X_1, X_2, ...$ как угодно зависимых случайных величин имело место

$$\frac{S_N}{N} \xrightarrow{p} \frac{1}{N}\sum_{i=1}^{N} E(X_i),$$

необходимо и достаточно, чтобы

$$\lim_{N\to\infty} E\left\{\frac{\left[\sum_{i=1}^{N}(x_i - E(X_i)\right]^2}{N^2+\left[\sum_{i=1}^{N}(x_i - E(X_i)\right]^2}\right\}=0.$$

Дальнейшие исследования показали, что статистики аддитивного типа не только сходятся к соответствующим средним по вероятности, как это утверждается законами больших чисел, но, начиная с некоторого N_0, со сколь угодно близкой к единице вероятностью S_N не выходит за пределы, сколь угодно близкие к среднему. Серия таких, более сильных, чем законы больших чисел, утверждений называется усиленными законами больших чисел.

Теорема 6.3.6. (Теорема Колмогорова). Если последовательность взаимно независимых случайных величин $X_1, X_2, \ldots$ удовлетворяет условию

$$\frac{1}{N^2}\sum_{i=1}^{N} D(X_i) < +\infty,$$

то S_N подчиняется усиленному закону больших чисел.

Теорема 6.3.7. (Теорема Колмогорова). Если $X_1, X_2, \ldots$ — последовательность независимых и одинаково распределенных случайных величин, то существование математического ожидания $E(X)$ является необходимым и достаточным условием для подчинения S_N усиленному закону больших чисел.

Если на переменные x_i наложить дополнительное требование ограниченности $x_i \leqslant a$, то усиленный закон больших чисел может быть уточнен.

Теорема 6.3.8. (Закон повторного логарифма). Пусть $X_1, X_2, \ldots$ — ограниченные независимые случайные переменные с одинаковыми средним $M=E(X_i)$ и дисперсией $D(X_i)=\sigma^2$. Обозначим $\eta_N=(S_N-NM)/\sigma$. Тогда

$$P\left(\limsup_{N\to\infty}\frac{|\eta_N|}{\sqrt{2N\ln\ln N}}\leqslant 1\right)=1.$$

В частности, справедлива

Теорема 6.3.9. Если $\frac{m}{N}$ — относительная частота события A, полученная в серии N независимых испытаний, а вероятность $P(A)$ события A равна $p\,(0<p<1,\ q=1-p)$, то

$$P\left(\limsup_{N\to\infty}\frac{\left|\frac{m}{N}-p\right|}{\sqrt{\frac{2pq}{N}\ln\ln N}}\leqslant 1\right)=1.$$

Интересно заметить, что хотя законы больших чисел характеризуют непараметрические свойства статистик, можно предположить, что они являются следствием каких-то непараметрических свойств самой выборки. На характер этих свойств выборки указывает теорема Шэннона-Макмиллана, которая утверждает, что все высоковероятные реализации выборки из дискретного распределения асимптотически равновероятны и что характеристическим параметром ансамбля этих реализаций является энтропия. (См., например, Ф. П. Тарасенко [4]).

§ 6.4. ЦЕНТРАЛЬНЫЕ ПРЕДЕЛЬНЫЕ ТЕОРЕМЫ ДЛЯ СУММ НЕЗАВИСИМЫХ СЛУЧАЙНЫХ ВЕЛИЧИН

Всякое статистическое решение принимается на основе некоторой статистики $S_N=S(x_1, \ldots, x_N)$. Для количественной характеристики получаемых решений необходимо знание распределения случайной величины S_N. Во многих случаях вычисление этого распределения даже при известном распределении выборочных значений является весьма сложным. Однако, оказывается, что при выполнении определенных условий распределение статистики S_N при возрастании N стремится к некоторому простому и известному распределению, вне зависимости от того, каков конкретный вид распределения выборочных значений. Такие утверждения называются предельными теоремами.

Вслед за классической теоремой Муавра-Лапласа целый цикл теорем, объединяемых названием центральных предельных теорем, определяет условия, при которых статистики аддитивного типа обладают асимптотическим нормальным распределением. В данном параграфе мы ограничимся перечислением основных предельных теорем для сумм независимых случайных величин.

Будем пользоваться обозначением

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{u^2}{2}}du. \qquad (6.4.1)$$

Начнем со следующего обобщения теоремы Муавра-Лапласа:

Теорема 6.4.1. (Теорема Линдеберга-Леви). Если X_1, X_2, ..., X_N — независимые и одинаково распределенные случайные величины, имеющие конечные среднее $E(X)$ и дисперсию $D(X)=\sigma^2$, и $S_N=\sum_i^N x_i$, то величина $t_N=(\frac{S_N}{N}-E(X))/\sigma$ имеет предельную функцию распределения $\Phi(t)$.

Если сделать предположение о существовании третьего момента, то одинаковость распределения для всех X_i является несущественной:

Теорема 6.4.2. (Теорема Ляпунова). Если X_1, X_2, ..., X_N — независимые случайные величины, имеющие конечные среднее $E(X_i)$, дисперсию σ_i^2 и третий абсолютный центральный момент $\rho_i^3=E\{|x_i-E(X_i)|^3\}$, и если выполняется условие $\lim_{N\to\infty}\rho/\sigma=0$, где $\rho^3=\sum_i^N\rho_i^3$, и $\sigma^2=\sum_i^N\sigma_i^2$, то величина $t_N=[S_N-\Sigma E(X_i)]/\sigma$ имеет предельное распределение $\Phi(t)$.

Центральная предельная теорема может быть доказана при еще более общих условиях:

Теорема 6.4.3. (Теорема Линдеберга). Пусть X_1, X_2, ..., X_N — независимые случайные переменные, имеющие конечные средние $E(X_i)$ и дисперсии $D(X_i)=\sigma_i^2$. Пусть $\sigma^2=\sum_{i=1}^N\sigma_i^2$ и $F_i(x)$ — функция распределения величины x_i. Если для любого $\varepsilon>0$ выполняется условие Линдеберга

$$\lim_{N\to\infty}\frac{1}{\sigma^2}\sum_{i=1}^{N}\int_{|x-E(X_i)|\geq\varepsilon\sigma}(x-E(X_i))\,dF_i(x)=0,$$

то для величины $t_N=[\sum_{i=1}^N(x_i-E(X_i)]/\sigma$ предельной функцией распределения является $\Phi(t)$.

Из теорем 6.4.1 — 6.4.3 следует, что для независимых случайных величин имеет место

$$\lim_{N\to\infty}P\left[\frac{S_N-A_N}{B_N}\leqslant t\right]=\Phi(t), \qquad (6.4.2)$$

где A_N определенным образом связано со средними значениями, а B_N — со стандартными отклонениями. Оказывается, что существование дисперсии может быть заменено более слабым требованием (6.4.3), если выбрать подходящим образом неслучайные последовательности чисел $\{A_N\}$ и $\{B_N\}$.

Теорема 6.4.4. (Теорема Леви-Феллера-Хинчина). Пусть X_1, X_2, ..., X_N — независимые случайные величины с одинаковой функцией распределения $F(x)$. Если для $F(x)$ выполняется предельное соотношение

$$\lim_{y\to\infty}\frac{y^2[F(-y)+(1-F(y))]}{\int\limits_{|x|<y}x^2dF(x)}=0, \qquad (6.4.3)$$

то (6.4.2) имеет место для любой, подходящим образом выбранной последовательности чисел $\{A_N\}$ и $\{B_N\}$.

Теоремы 6.4.1 — 6.4.4 утверждают сходимость функции распределения стандартизованных сумм t_N случайных величин X_t к интегральному нормальному распределению. Естественно возникает вопрос, при каких условиях плотность вероятностей величины t_N сходится к плотности нормального распределения. Очевидно, эти условия будут более строгими, чем для теорем 1—4, так как из сходимости $F_N(t)$ к $\Phi(t)$ еще не следует $F'_N(t)\to\Phi'(t)$. Этот вопрос был детально изучен Б. В. Гнеденко; основной результат можно сформулировать следующим образом:

Теорема 6.4.5. (Локальная предельная теорема Гнеденко). Пусть $X_1, X_2, \ldots, X_N$ — независимые, одинаково распределенные случайные величины, имеющие ограниченную плотность $p(x)$. Предположим далее, что $E(X_t)=\int\limits_{-\infty}^{\infty}xp(x)dx=0$ и что интеграл $\sigma^2=\int\limits_{-\infty}^{\infty}x^2p(x)dx$ существует. Тогда плотность $f_N(t)$ вероятностей величины

$$t_N=\frac{x_1+x_2+\ldots+x_N}{\sigma\sqrt{N}}$$

равномерно по x стремится к плотности нормального распределения:

$$\lim_{N\to\infty}f_N(t)=\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}.$$

§ 6.5. ЦЕНТРАЛЬНЫЕ ПРЕДЕЛЬНЫЕ ТЕОРЕМЫ ДЛЯ СУММ ЗАВИСИМЫХ СЛУЧАЙНЫХ ВЕЛИЧИН

Во многих приложениях предположение о независимости выборочных значений выполняется лишь приближенно, а иногда и заведомо не выполняется. Можно ожидать, что утверждение о предельной нормальности сумм случайных величин останется в силе, если зависимость между слагаемыми будет «достаточно слабой» или если она сильна для соседних выборочных значений, но убывает «достаточно быстро» при увеличении разности их номеров. В последние годы различные исследователи много внимания уделяют выяснению того, какие же зависимости являются достаточно слабыми

или достаточно быстро убывающими для того, чтобы можно было пользоваться нормальной аппроксимацией предельных законов распределения. Приведем здесь некоторые из результатов в этом направлении, ориентированные на непараметрическую статистику.

Примером последовательности случайных величин, зависимость между которыми не связана с их номерами в выборке, но является достаточно слабой при достаточно больших объемах выборки, могут служить компоненты $R_1, R_2, \ldots, R_N$ рангового вектора при равновероятности перестановок выборочных значений. В этом случае может быть доказана асимптотическая нормальность линейных ранговых статистик*.

Дадим соответствующую теорему Вальда-Вольфовица-Нётера-Гаека в формулировке Я. Гаека [1]:

Теорема 6.5.1. Пусть

$$S_N = \sum_{i=1}^{N} C_i \left[u + v\varphi\left(\frac{R_i}{N+1} \right) \right],$$

где $v \neq 0$ и φ — либо неубывающая на $[0, 1]$ функция, либо функция, невозрастающая на $[0, a]$ и неубывающая на $[a, 1]$, для некоторого $a \in (0, 1)$.

Предположим, что

$$0 < \int_0^1 [\varphi(t) - \bar{\varphi}]^2 \, dt < \infty, \quad \bar{\varphi} = \int_0^1 \varphi(t) \, dt.$$

Тогда для любого $\varepsilon > 0$ существует такое $\delta = \delta(\varepsilon, \varphi)$, что при выполнении условия

$$\max_{1 \leqslant i \leqslant N} \{(C_i - \bar{C})^2\} < \delta \sum_{i=1}^{N} (C_i - \bar{C})^2$$

имеет место сходимость распределения S_N к нормальному:

$$\sup_{-\infty < s < \infty} \left| P(S_N \leqslant s) - \Phi\left[\frac{s - E(S_N)}{\sigma_N} \right] \right| < \varepsilon.$$

Обратимся теперь к случаю, когда зависимость между двумя выборочными значениями X_t и X_k может быть сильной

---

* Линейными ранговыми статистиками называются статистики вида $S_N = \sum_i C_i a(R_i)$, где $\{C\}$ — так называемые регрессионные константы, а $a(i)$ — весовые коэффициенты. Говорят, что веса $a(i)$ порождаются функцией $\varphi(x)$, если $a(i) = u + v\varphi\left(\frac{i}{N+1}\right)$, где u и v — произвольные константы.

при малых значениях $|i-k|$, но быстро ослабевает при увеличении $|i-k|$.

Удалось доказать центральную предельную теорему для так называемых m-зависимых процессов. Последовательность случайных переменных $X_1, X_2, \ldots$ называется m-зависимой, если при $s-r>m$ $(X_1, \ldots, X_r)$ и $(X_s, X_{s+1}, \ldots)$ независимы. Иначе говоря, если из последовательности $X_1, X_2, \ldots$ изъять m или больше соседних наблюдений, то оставшиеся две подпоследовательности будут независимыми. Обозначим

$$A_i=2\sum_{j=0}^{m-1}\operatorname{cov}\{X_{i\ j},\ X_{i+m}\}-\operatorname{var}\{X_{i+m}\}.$$

Теорема 6.5.2. (Теорема Хёфдинга-Робинса). Если

а) m-зависимая последовательность $X_1, X_2, \ldots$ удовлетворяет соотношениям $E(X_i)=0$, $E(|X_i|^3)\leqslant\rho^3<\infty$ для всех i,

б) $\lim\limits_{k\to\infty}\dfrac{1}{k}\sum\limits_{h=1}^{k}A_{i+h}=A$ для всех i,

то $S_N=\sum\limits_{i=1}^{N}X_i$ асимптотически нормальна со средним нуль и дисперсией NA.

Для стационарных m-последовательностей теорема 6.5.2 упрощается:

Теорема 6.5.3. Если $X_1, X_2, \ldots$ — стационарная m-зависимая последовательность случайных величин с существующими $E(X_1)=\mu$ и $E(|X_i|^3)$, то предельное распределение величины $S_N/\sqrt{N}$ нормально со средним $\sqrt{N}\mu$ и дисперсией A,

$$A=\operatorname{var}(X_1^2)+2[\operatorname{cov}(X_1X_2)+\ldots+\operatorname{cov}(X_1X_{m+1})].$$

Другой пример случайных последовательностей зависимых случайных величин, для которых удается доказать центральную предельную теорему, дают марковские процессы. Правда, строгие доказательства удалось построить пока для случая однородных цепей Маркова с конечным числом состояний (см. А. Реньи [1]).

§ 6.6. О ПРЕДЕЛЬНЫХ РАСПРЕДЕЛЕНИЯХ НЕ НОРМАЛЬНОГО ТИПА

До сих пор мы обсуждали условия, при которых распределения статистик аддитивнго типа сходятся к нормальному распределению. Возникает вопрос, существуют ли другие предельные распределения для сумм независимых случайных величин? Оказалось, что нормальное распределение является не единственно возможным предельным законом.

Можно показать, что класс предельных распределений для сумм независимых случайных величин совпадает с классом так называемых безгранично-делимых законов*. Наиболее интересным для нас результатом теории безгранично-делимых законов является

Теорема 6.6.1. (Теорема Гнеденко). Всякий безгранично-делимый закон является либо композицией конечного числа законов Пуассона и нормального закона, либо пределом равномерно сходящейся последовательности таких законов.

Таким образом, любое предельное распределение суммы независимых случайных величин как бы составлено из пуассоновского и нормального законов. К числу безгранично-делимых законов относятся распределение Паскаля

$$P(X=k)=\frac{a^k}{(1+a)^{k+1}},\ a>0,\ k=1,\ 2,\ \ldots, \tag{6.6.1}$$

распределение Пойа

$$P(X=k)=\left(\frac{\alpha^\lambda}{1+\alpha^\lambda}\right)^k\cdot\frac{1}{k!}\cdot(1+\alpha)\ldots(1+(k+1)\alpha)\cdot p_0, \tag{6.6.2}$$

$$k=1,\ 2,\ldots,\ \alpha>0;\ \lambda>0;\ p_0=P(X=0)=(1+\alpha^\lambda)^{-\frac{1}{\alpha}};$$

распределение Коши

$$p(x)=\frac{ab}{\pi(a^2+b^2x^2)},\ -\infty<x<\infty; \tag{6.6.3}$$

гамма-распределение

$$p(x)=\frac{\beta^m}{\Gamma(m)}e^{m-1}e^{-\beta x},\ x\geqslant 0; \tag{6.6.4}$$

χ^2 — распределение и другие.

Хотя статистики аддитивного типа являются важным классом статистик, они, конечно, не исчерпывают всех статистик, применяемых для вынесения решений. Поэтому естественны поиски предельных теорем для других типов статистик. Для ряда статистик эти поиски оказались успешными; например, для таких статистик неаддитивной структуры, как порядковые статистики и выборочные интервалы, соответствующие результаты приведены в главах III и IV. Известны и другие результаты, но систематической теории пока не построено.

---

* Закон $F(x)$ называется безгранично-делимым, если при любом N его характеристическая функция является N-й степенью некоторой другой характеристической функции.

ДОБАВЛЕНИЕ

О НЕКОТОРЫХ ОБЩИХ СВОЙСТВАХ РАСПРЕДЕЛЕНИЙ

§ Д.1. ВВЕДЕНИЕ

В предыдущих главах данной Части наше внимание было сосредоточено на отыскании и рассмотрении непараметрических свойств различных статистик (т. е. функций выборочных значений): порядковых статистик (гл. III), выборочных интервалов (гл. IV), рангов (гл. V), статистик аддитивной структуры (гл. VI). Каждый раз мы убеждались, что все эти статистики обладают некоторыми свойствами, не зависящими от конкретного вида распределения, из которого взята выборка (правда, в ряде случаев такие свойства проявляются лишь асимптотически). Никакая обработка выборочных значений не может дать информации, которая не содержалась бы в этих значениях. Раз уж мы обнаруживаем непараметрические свойства у статистик, значит таковыми обладает и сама выборка, а следовательно, и ее распределение. Способ обработки выборки (построение статистик) есть лишь преобразование, позволяющее извлечь нужную нам информацию, представить ее в удобном для использования виде. Очевидно, информация, необходимая нам для решения статистических задач, есть информация о различиях между распределениями. В непараметрических задачах речь может идти только о таких различиях, которые не связаны с конкретным функциональным видом распределений, т. е. о различиях между классами, типами, семействами распределений. Естественно попытаться как-то конкретизировать характер таких различий. Трудность здесь заключается в том, что говорить о «различиях вообще» кажется малоперспективным: для одних задач данное различие является существенным, другие задачи являются по отношению к нему инвариантными. Тем не менее, мы обсудим некоторые возможные типы различий между распределениями безотносительно к задачам, лишь иногда вспоминая, что каждому из этих типов соответствует определенный класс задач.

§ Д.2. НЕКОТОРЫЕ НЕПАРАМЕТРИЧЕСКИЕ КЛАССЫ ОДНОМЕРНЫХ РАСПРЕДЕЛЕНИЙ

Первым основанием классификации распределений само собой напрашивается разделение их на симметричные и

несимметричные. Формально распределение называется симметричным относительно точки μ, если его функция распределения $G(x)$ удовлетворяет соотношению

$$G(-x-\mu)=1-G(x-\mu) \quad \text{для всех } x\in X, \tag{Д.2.1}$$

а в случае существования плотности $g(x)$ —

$$g(-x-\mu)=g(x-\mu) \quad \text{для всех } x\in X. \tag{Д.2.2}$$

Отметим, что задание точки μ обязательно: распределение, симметричное относительно μ, не будет симметричным относительно любой другой точки $\mu_1\neq\mu$.

Следующим, иногда полезным признаком классификации является разбиение на классы по числу максимумов (супремумов) у плотности. В случае одного максимума распределение называется одновершинным (или одномодовым); точка из X, соответствующая положению максимума, называется модой. Далее, если функция $-\log g(x)$ выпукла в некотором интервале (a, b), таком, что $-\infty\leqslant a<b\leqslant\infty$ и $\int_a^b g(x)dx=1$, то плотность $g(x)$ называется усиленно одновершинной (Я. Гаек и З. Шидак [1]). Все распределения, описанные в следующем параграфе, удовлетворяют условию усиленной одновершинности.

Обобщение указанных выше двух классификаций на многомерные распределения представляется очевидным.

§ Д.3. ТИПЫ РАСПРЕДЕЛЕНИЙ

Будем говорить, что распределения $F(x)$ и $G(x)$ принадлежат к одному типу, если для всех $x\in X$

$$F(x)=G\left(\frac{x-\mu}{\sigma}\right), \tag{Д.3.1}$$

где $\mu\in(-\infty, \infty)$, $\sigma>0$. При существовании плотностей они принадлежат к одному типу, если для всех $x\in X$

$$f(x)=\frac{1}{\sigma}\,g\left(\frac{x-\mu}{\sigma}\right). \tag{Д.3.2}$$

Константы μ и σ называются параметрами расположения (сдвига) и масштаба соответственно. Таким образом, любое распределение данного типа может быть получено из другого распределения того же типа с помощью линейного преобразования аргумента.

Класс однотипных распределений не обязательно является параметрическим семейством — в том смысле, что мы можем

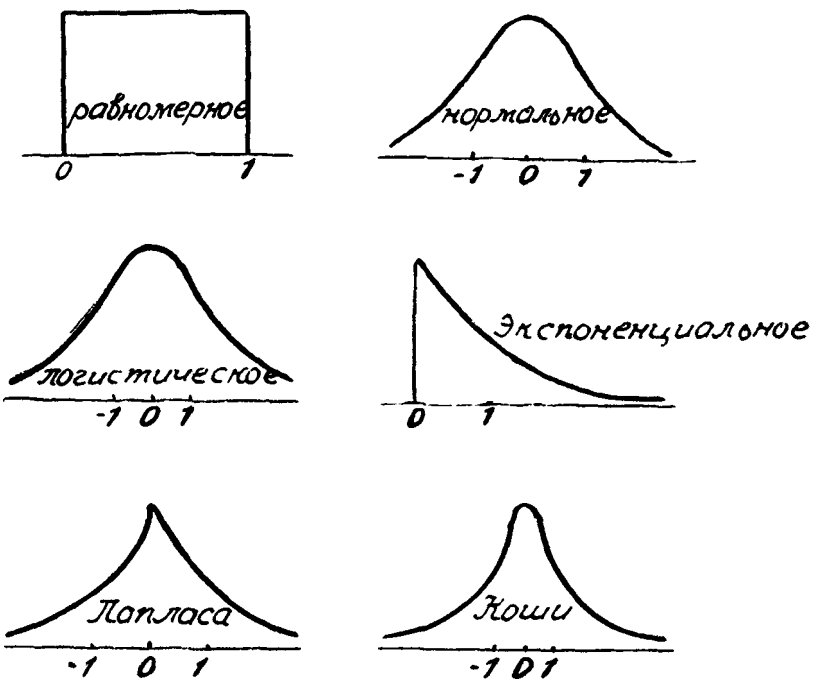

Рис Д 3 1

и не знать конкретного вида функций F и G. Конечно, если задать некоторое распределение, то класс распределений данного типа будет параметрическим. Так как численные характеристики качества статистических процедур могут быть получены лишь для конкретных типов распределений, то полезно перечислить наиболее употребительные из распределений. Ниже мы приводим такой список; графики, представляющие вид соответствующих плотностей, приведены на рис. Д.3.1.

1. Равномерная в $[0, 1]$ плотность:

$$f(x)=\begin{cases}1, & x\in[0,1],\\ 0, & x\notin[0,1].\end{cases}$$

2. Нормальная плотность:

$$f(x)=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{x^2}{2}\right],\quad -\infty\leqslant x\leqslant\infty.$$

3. Логистическая плотность:

$$f(x)=\frac{e^{-x}}{(1+e^{-x})^2},\quad -\infty\leqslant x\leqslant\infty.$$

4. Экспоненциальная плотность:

$$f(x)=e^{-x},\ x\geqslant 0.$$

5. Двойное экспоненциальное распределение (Лапласа)

$$f(x)=\frac{1}{2}e^{-|x|},\quad -\infty\leqslant x\leqslant\infty.$$

6. Плотность Коши:

$$f(x)=\frac{1}{\pi}\cdot\frac{1}{1+x^2},\ -\infty\leqslant x\leqslant\infty.$$

§ Д.4. УПОРЯДОЧИВАНИЕ СИММЕТРИЧНЫХ РАСПРЕДЕЛЕНИЙ ПО ЗАТЯНУТОСТИ ХВОСТОВ

Многочисленность непараметрических процедур для решения одной и той же задачи выдвигает проблему их сравнения между собой. Такое сравнение можно провести, например, рассмотрев, насколько устойчивыми окажутся те или иные количественные характеристики процедур при переходе от одного типа распределения выборки к другому.

Общность такого сравнения, его целенаправленность резко возросли бы, а необходимое количество перебираемых при этом типов распределений уменьшилось, если бы удалось упорядочить типы распределений по некоторым конкретным признакам различия между ними. Вопрос о существовании таких признаков отпал: Ван-Цвету [1] и Гаеку [1] удалось показать, что для симметричных распределений по крайней мере один такой признак существует; интригующим является вопрос, существуют ли другие признаки различия, допускающие упорядочивание типов распределений.

Первым необходимым шагом для выявления искомых признаков является стандартизация параметров сдвига и масштаба распределений сравниваемых типов. Естественными и всегда существующими (в отличие, например, от моментов) характеристиками сдвига и масштаба являются квантили и интерквантильные интервалы. Для симметричных распределений достаточно задание всего двух квантилей. Сдвиг у сравниваемых распределений будет одинаковым, если мы потребуем одинаковости их медиан (которые без потери общности можно положить равными нулю), а масштаб симметричных распределений автоматически стандартизуется, если потребовать одинаковость квантилей некоторого уровня, отличного от 0,5. Для определенности потребуем, кроме равенства медиан, равенства 95% квантилей. При этом вместо исходных плотностей $f(x)$, приведенных в § Д.3, необходимо

перейти к плотностям $g(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$. Далее, зададим μ и σ так, что

$$F\left(-\frac{\mu}{\sigma}\right) = 0{,}5 \text{ и } F\left(\frac{1{,}645-\mu}{\sigma}\right) = 0{,}95. \qquad (Д.4.1)$$

При таком определении μ и σ все сравниваемые распределения будут иметь те же медиану и 95%-ный квантиль, что и стандартное нормальное распределение. При этом перечисленные в § Д.3 плотности будут иметь параметры μ и σ, приведенные ниже в таблице.

Тип распределения	μ	σ
Равномерное	—1,827	3,655
Нормальное	0	1
Логистическое	0	0,560
Лапласа	0	0,714
Коши	0	0,261

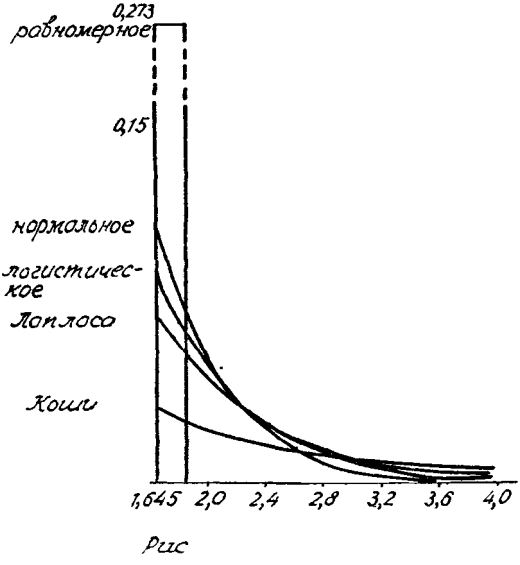

Рис Д.4.1

Оказывается, что после такого приведения распределений к одинаковым сдвигу и масштабу все типы распределений упорядочиваются по затянутости их хвостов (см. рис. Д.4.1). Самые короткие хвосты имеет равномерное распределение, затем по порядку идут нормальное, логистиче-

ское, двойное экспоненциальное; распределение Коши имеет наиболее затянутые хвосты. Это качество распределений можно описать и аналитически. Говорят (Гаек [1]), что $F(x)$ имеет более затянутые хвосты, чем $G(x)$, если

$$F^{-1}(u)=a(u)\,G^{-1}(u),\ \frac{1}{2}<u\leqslant 1, \qquad (Д.4.2)$$

где $a(u)$ — неубывающая функция.

Может показаться, что упорядочивание типов распределений по затянутости хвостов носит частный характер, описывает свойства только самих хвостов. Покажем, что такое упорядочивание затрагивает свойства распределений в целом.

Теорема Д.4.1. Все распределения F_i, входящие в класс непрерывных и симметричных, имеющих одновершинные плотности f_i и равные медианы (которые без потери общности положим равными нулю), при любых $u\in\left(\frac{1}{2},\ 1\right)$ однозначно упорядочиваются по значениям плотностей в точке x_u, соответствующей квантилю заданного уровня u, если потребовать, чтобы для всех f_i x_u было одним и тем же:

$$\ldots>f_1(F_1^{-1}(u))\geqslant f_2(F_2^{-1}(u))>\ldots>f_k(F_k^{-1}(u))>\ldots, \qquad (Д.4.3)$$

где равенства выполняются, если и только если соответствующие плотности принадлежат одному типу.

(Последнее утверждение является следствием Леммы Вайсса (см. ниже)).

Доказательство. Меньшая затянутость хвостов у F_i по сравнению с F_j означает, что, начиная с некоторого $x_0(F_i,\ F_j)$, для любых $x>x_0$ имеет место $f_i(x)<f_j(x)$. Поэтому из монотонности f_i и f_j при $x>0$ и равенства $\int\limits_{x_u}^{\infty} f_i(x)\,dx=\int\limits_{x_u}^{\infty} f_j(x)\,dx$ (т. е. $F_i[F_i^{-1}(u)]=F_j[F_j^{-1}(u)]=u$) следует $f_j[F_j^{-1}(u)]>f_i[F_i^{-1}(u)]$, что и требовалось доказать.

Отметим, что упорядоченность (Д.4.3) сохраняется при любых u, т. е. свойство, названное «затянутостью хвостов», на самом деле есть свойство распределений в целом.

Лемма Вайсса Д.4.2. Пусть $F(x)$ и $G(x)$ — две функции распределения, u и v $(0\leqslant u<v\leqslant 1)$ — два заданных числа. Предположим, что $F^{-1}(u)$, $F^{-1}(v)$, $G^{-1}(u)$ и $G^{-1}(v)$ все определены однозначно. Пусть $F(x)$ имеет производную $f(x)$ на $[F^{-1}(u),\ F^{-1}(v)]$, а $G(x)$ — производную $g(x)$ на $[G^{-1}(u),\ G^{-1}(v)]$. Если $F(x)=G\left(\frac{x-\mu}{\sigma}\right)$ для всех x из интервала $[F^{-1}(u),\ F^{-1}(v)]$, при некоторых μ и σ $(\sigma>0)$, то для всех $y\in[u,\ v]$

$$f\left(F^{-1}(y)\right)=\frac{1}{\sigma}\,g\left(G^{-1}(y)\right). \qquad (Д.4.4)$$

Кроме того, если $f(x)>0$ для $x\in[F^{-1}(u), F^{-1}(v)]$, то верно обратное утверждение.

Доказательство дано Вайссом [4]. Блюменталь [3] указал на пример, показывающий необходимость дополнительного условия для обратного утверждения.

Результаты данного параграфа будут использованы в дальнейшем, в частности, в главе IX.

§ Д.5. КАНОНИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ ПАР РАСПРЕДЕЛЕНИЙ

Даже если тест рассчитан на проверку сложных гипотез, многие суждения о его свойствах могут быть сделаны при задании конкретных пар гипотетического и альтернативного распределений и совокупностей таких пар. Во многих случаях полезно пользоваться некоторым унифицированным представлением пар распределений. Одно из таких представлений (называемое каноническим) мы рассмотрим в данном параграфе.

Пусть заданы распределение $F(x)$, соответствующее гипотезе, и альтернативное распределение $G(x)$. Выборке одинаково распределенных и независимых величин $x_1, \ldots, x_N$ поставим в соответствие совокупность чисел $F(x_1), \ldots, F(x_N)$. Будем называть эту совокупность приведенной выборкой, функцию $F(x)$ — функцией приведения, а совокупность возможных значений случайной величины $U=F(X)$ (интервал $[0, 1]$) — приведенным пространством, в отличие от исходного пространства X. Легко показать, что если выборка взята из распределения $G(x)$, то функция распределения приведенной выборки выразится как

$$\tau(u)=G[F^{-1}(u)],\ u\in[0,\ 1], \qquad (Д.5.1)$$

а плотность запишется в виде:

$$g^*(u)=\frac{g[F^{-1}(u)]}{f[F^{-1}(u)]},\ u\in[0,1], \qquad (Д.5.2)$$

где f и g — плотности соответственно для F и G. Будем называть (Д.5.1) и (Д.5.2) канонической формой альтернативы. Очевидно, гипотеза в канонической форме всегда есть равномерное в $[0, 1]$ распределение.

Полезность канонического представления распределений можно продемонстрировать на ряде примеров. Предположим, что мы рассматриваем пары распределений F и G одного

типа, отличающиеся только сдвигом, т. е. $G(x)=F(x-\mu)$. Если теперь зафиксировать сдвиг μ и перебрать известные нам типы распределений, то обнаружится, что канонические альтернативы упорядочатся по степени отличия от канонической гипотезы точно в том же порядке, в каком исходные альтернативы упорядочены по хвостам (см. рис. Д.5.1). И на-

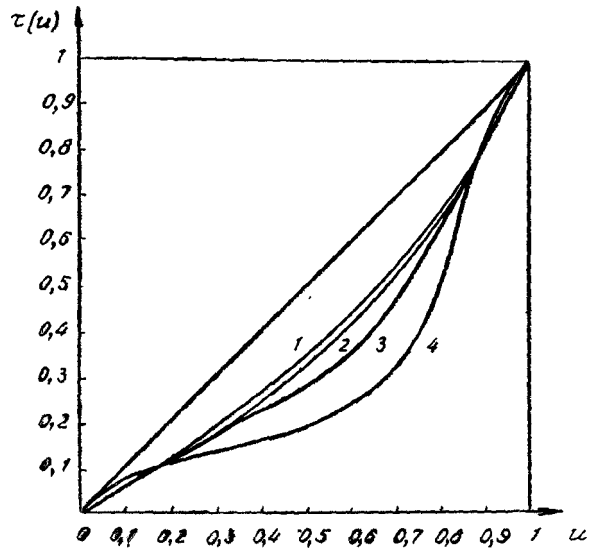

Рис. Д.5.1

оборот, если рассматривать масштабные альтернативы вида $G(x)=F(\lambda x)$, то при одном и том же λ канонические альтернативы будут тем сильнее отличаться от гипотезы, чем менее затянут хвост распределения (рис. Д.5.2). Эти примеры еще раз показывают, что классификация по затянутости хвостов затрагивает такие различия между распределениями, которые существенным образом влияют на свойства тестов. Для удобства читателя приведем аналитические выражения канонических альтернатив, графики которых представлены на рис. Д.5.1 и Д. 5.2.

Альтернативы сдвига ($G(x)=F(x-a)$). Значения σ соответствуют данным таблицы в § Д.4.

Нормальная

$$\tau_2(u/a)=\Phi\left[\Phi^{-1}(u)-a\right],\ u\in[0,\ 1]. \qquad (Д.5.3)$$

Логистическая

$$\tau_3(u/a)=\left[1-e^{-\frac{a}{\sigma}}\left(\frac{1}{u}-1\right)\right]^{-1},\ u\in[0,1]. \qquad (Д.5.4)$$

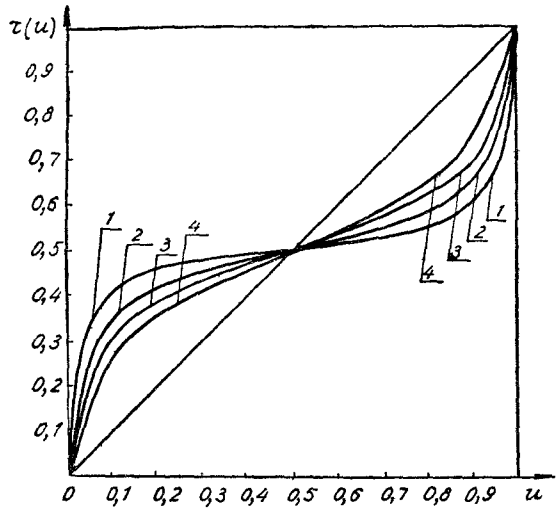

Рис. Д.5.2

Лапласа

$$\tau_5(u/a)=\begin{cases} ue^{-\frac{a}{\sigma}}, & 0\leqslant u<\dfrac{1}{2}, \\ \dfrac{1}{4}e^{-\frac{a}{\sigma}}(1-u)^{-1}, & \dfrac{1}{2}\leqslant u\leqslant 1-\dfrac{1}{2}e^{-\frac{a}{\sigma}}, \\ 1-e^{-a}(1-u), & 1-\dfrac{1}{2}e^{-\frac{a}{\sigma}}\leqslant u\leqslant 1. \end{cases} \quad (Д.5.6)$$

Коши:

$$\tau_6(u/a)=\frac{1}{2}+\frac{1}{\pi}\operatorname{arctg}\left\{\operatorname{tg}\left[\pi\left(u-\frac{1}{2}\right)\right]-\frac{a}{\sigma}\right\},\ u\in[0,1]. \quad (Д.5.7)$$

Альтернативы масштаба: $\left(G(x)=F\left(\dfrac{x}{\theta}\right)\right)$:

Нормальная: $\tau_2(u/\theta)=\Phi[\Phi^{-1}(u)\cdot\theta],\ u\in[0,1].$ (Д.5.8)

Логистическая:

$$\tau_3(u/\theta)=\left[1+\left(\frac{1}{u}-1\right)^{\theta}\right]^{-1},\ u\in[0,1]. \quad (Д.5.9)$$

Лапласа:

$$\tau_5(u/\theta)=\begin{cases} \dfrac{1}{2}[2u]^{\theta}, & 0\leqslant u\leqslant\dfrac{1}{2}, \\ 1-\dfrac{1}{2}[2(1-u)]^{\theta}, & \dfrac{1}{2}\leqslant u\leqslant 1. \end{cases} \quad (Д.5.10)$$

Коши:

$$\tau_6(u, \theta)=\frac{1}{2}+\frac{1}{\pi}\operatorname{arctg}\left\{\theta\cdot\operatorname{tg}\left[\pi\left(u-\frac{1}{2}\right)\right]\right\},\quad u\in[0,\ 1]. \qquad (Д.5.11)$$

§ Д.6. ОБ ОДНОМ СВОЙСТВЕ КАНОНИЧЕСКОГО ПРЕДСТАВЛЕНИЯ ПАР РАСПРЕДЕЛЕНИЙ

Исследование тестов с помощью рассмотрения изменений их свойств при задании различных пар «гипотеза — альтернатива», в свою очередь, выдвигает вопрос о некотором принципе упорядочивания таких пар между собой. В качестве такого принципа следует выбрать некоторым разумным образом определенную меру различия между альтернативой F и гипотезой G, так как надежность формируемых тестом решений, как правило, возрастает при увеличении различий между F и G. Минимум требований, которые естественно предъявить к любой разумной мере расхождения между альтернативой и гипотезой, сводится к двум условиям:

1) мера должна быть инвариантной по отношению к любым взаимно-однозначным преобразованиям случайной величины X (при этом подразумевается, что если X подвергнуть преобразованию, то F и G преобразуются одновременно);

2) если рассматривается совокупность независимых случайных величин $\{X_i\}$, с каждой из которых связана пара распределений F_i и G_i, то вводимая мера должна обладать свойством аддитивности.

Можно сделать вполне определенные суждения о том, какого вида функционалы обладают указанными свойствами. Пусть $I(f, g)$ есть функционал, выражающий искомую меру. Если теперь подвергнуть случайную величину X каноническому преобразованию $y=F(x)$, то в силу Д.5.2 и требования инвариантности I имеем:

$$I(f, g)=I\left(1, \frac{g}{f}\right), \qquad (Д.6.1)$$

откуда следует, что

$$I(f, g)=\int\varphi\left(\frac{g}{f}\right)d\mu, \qquad (Д.6.2)$$

где φ есть некоторая нелинейная функция своего аргумента, а μ — вероятностная мера, связанная с F и/или G, так как I должно характеризовать только эту пару распределений. Поскольку мы собираемся связывать рассмотрение этого «расстояния» с мощностными свойствами тестов, определяемыми

G, то в качестве μ целесообразно взять именно альтернативное распределение G. Далее, требование аддитивности вводимой меры приводит к единственно допустимой функции φ, а именно — к логарифмической функции. Таким образом, мы останавливаемся на характеристической мере расстояния распределений, ориентированной на использование при анализе мощностных свойств тестов, в виде функционала вида:

$$I = \int \ln \frac{g}{f}\, dG, \qquad \text{(Д.6.3)}$$

который известен (Кульбак [1]) как средняя информация для различения в пользу G против F. Нетрудно показать (лемма 4.1 Кульбак [1]), что (Д.6.3) удовлетворяет требованию 1 не только при преобразовании $y=F(x)$, но и при любом взаимно-однозначном преобразовании. Отметим далее интересный факт, заключающийся в том, что информационное расстояние (Д.6.3) является негэнтропией канонически приведенной альтернативы. В самом деле, пользуясь формулой (Д.5.2), имеем:

$$-H(g^*) = \int_0^1 g^* \ln g^*\, du = \int \ln \frac{g}{f} \cdot dG. \qquad \text{(Д.6.4)}$$

В дальнейшем нам понадобятся зависимости энтропии приведенных альтернатив (Д.5.3)—(Д.5.11) от параметров сдвига и масштаба.

Альтернатива		Энтропия
Нормальная	сдвиг (Д.5.3)	$\frac{a^2}{2\sigma^2}$
	масштаб (Д.5.8)	$\frac{1}{2}(1-\theta^{-2})-\ln\theta$
Логистическая	сдвиг (Д.5.4)	$2+\frac{a}{\sigma}\cdot\frac{1+e^{\frac{a}{\sigma}}}{1-e^{\frac{a}{\sigma}}}$
	масштаб (Д.5.9)	—
Лапласа	сдвиг (Д.5.6)	$\frac{a}{2\sigma}(e^{-\frac{a}{\sigma}}+e^{\frac{a}{\sigma}})$
	масштаб (Д.5.10)	$\frac{1-\theta}{\theta}-\ln\theta$
Коши	сдвиг (Д.5.7)	—
	масштаб (Д.5.11)	$\ln\left[\theta\left(\frac{2}{1+\theta}\right)^2\right]$.

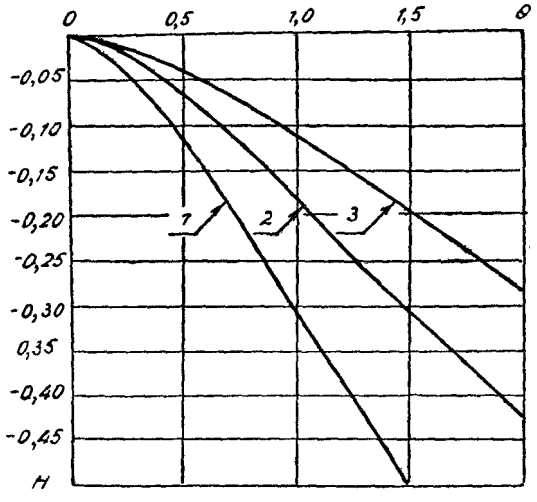

Рис. Д 6 1

Графики некоторых из приведенных выше функций для масштабных альтернатив даны на рис. Д.6.1. (*1* — нормальная, *2* — Лапласа, *3* — Коши).

Часть III

СТАТИСТИЧЕСКИЕ ПРОЦЕДУРЫ РЕШЕНИЯ НЕПАРАМЕТРИЧЕСКИХ ЗАДАЧ

ГЛАВА VII

ОЦЕНИВАНИЕ НЕИЗВЕСТНЫХ РАСПРЕДЕЛЕНИЙ ВЕРОЯТНОСТЕЙ

§ 7.1. ЗАДАЧА ОЦЕНИВАНИЯ РАСПРЕДЕЛЕНИЙ. ВИДЫ СХОДИМОСТИ ОЦЕНОК НЕПРЕРЫВНЫХ ФУНКЦИЙ

Во многих практических случаях возникает необходимость по экспериментальным данным оценить распределение вероятностей измеренной случайной величины. Эта задача имеет свою долгую историю, которая, судя по всему, еще не закончена. Было предложено и развито много различных подходов к этой задаче, каждый из которых обладает преимуществами перед другими в некоторой конкретной ситуации. Все эти подходы достаточно четко можно разделить на два класса — параметрическое и непараметрическое оценивание распределений.

Параметрический подход к оцениванию распределений сводится к тому, чтобы из некоторого, более или менее широкого параметрического класса распределений подобрать то, которое наилучшим образом соответствует выборке. Например, центральная предельная теорема гарантирует, что во многих (хотя далеко не всех) случаях приемлемой аппроксимацией может служить нормальное распределение; в этих случаях задача сводится к оценке по выборке среднего и дисперсии. Если заведомо известна ограниченность возможных значений случайной величины X с одной стороны, параметрический подход предлагает пользоваться семействами логнормальных или гамма-распределений; при ограниченности X сверху и снизу — семейством бета-распределений. Если ввести в рассмотрение такие показатели формы распределения, как квадрат нормированного коэффициента асимметрии (β_1) и нормированный показатель островершинности (β_2), выражаемые через моменты третьего и четвертого порядка, то можно указать области значений (β_1, β_2), в которых распределения принадлежат к тому или иному типу (см. график Пирсона, рис. 7.1.1, заимствованный из книги Г. Хана и С. Шапиро [1]). Существует, однако, область (заштрихованная на графике), которая не охватывается перечисленными

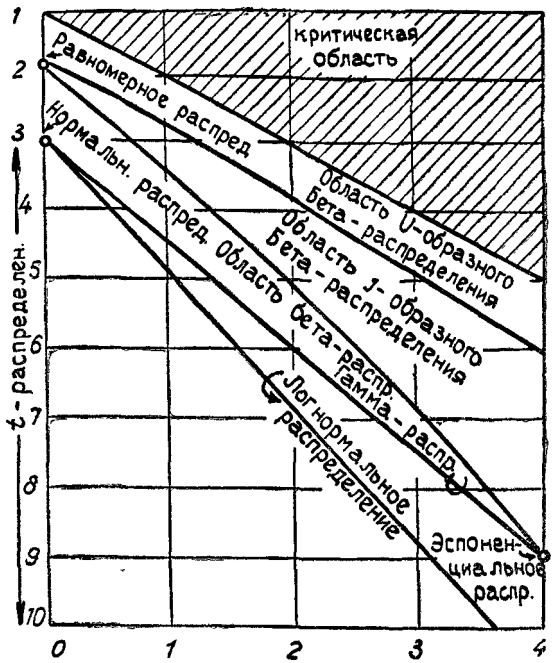

Рис. 7.1.1

выше распределениями. Чтобы перекрыть и ее, Джонсон [1] предложил рассматривать параметрические классы распределений, получаемые некоторыми трансцендентными преобразованиями нормальной случайной величины, а Пирсон (см. Кендалл и Стьюарт [1]) — класс распределений, порождаемых определенным дифференциальным уравнением (параметричность этого класса связана с коэффициентами уравнения). Другие типы параметрических оценок распределений получаются при использовании конечного числа членов разложений Корниша — Фишера, Грама — Шарлье, Эджворта и т. д. Описание упомянутых выше параметрических методов интересующийся читатель может найти, например, в книгах Кендалла и Стьюарта [1], Хана и Шапиро [1], А. К. Митропольского [1] и др. В данной главе мы сосредоточим внимание на *непараметрических* методах оценивания неизвестных распределений.

Прежде чем перейти к конкретным непараметрическим оценкам плотностей или функций распределения, обсудим вопрос о возможных типах сходимости статистических оценок непрерывных функций к оцениваемым функциям. Для

получения непараметрических оценок функции распределения F или плотности p могут быть использованы различные статистические факты, что приводит к оценкам различных типов и качества. Предпочтение обычно отдается оценкам, лучшим по качеству, однако в ряде практических ситуаций на первый план может выступить объем вычислений, необходимый для достижения результата (в особенности в многомерном случае).

Рассмотрим, какими свойствами могут обладать статистические оценки функций, и как их качества могут быть описаны количественно. Поскольку от оценки функции естественно ожидать определенного приближения, стремления к неизвестной оцениваемой функции, то прежде всего необходимо установить, какие типы сходимости одной функции к другой вообще возможны. Для простоты будем пока говорить об одномерных функциях.

Пусть имеется некоторая неизвестная функция $y(x)$ и ее статистическая оценка $y_N(x)=y(x|x_1, ..., x_N)$, построенная на основе выборки $x_1, ..., x_N$ с использованием определенной априорной информации. (В непараметрической постановке задачи об оценивании функции $y(x)$ роль априорной информации играет тот или иной непараметрический факт, имеющий отношение к функции $y(x)$).

Можно тем или иным образом определить «расстояние» между $y(x)$ и $y_N(x)$, т. е. числовую меру различия между ними, $\rho_N=\rho(y_N(x), y(x))$. Тогда сходимость последовательности функций $y_N(x)$ к оцениваемой функции $y(x)$ естественно понимать как сходимость последовательности случайных величин ρ_N к нулю.

В связи с тем, что, с одной стороны, к последовательности можно предъявить различные требования по сходимости (например, сходимость по вероятности, т. е. $\lim\limits_{N\to\infty} Pr(\rho_N\geqslant\varepsilon)=0$ или сходимость в среднем порядка r, т. е. $\lim\limits_{N\to\infty} E(|\rho_N{}^r|)=0$), а с другой — можно по-разному определить расстояние ρ_N о сходимости функций $y_N(x)$ к $y(x)$ можно говорить в различных смыслах. Так, если $\rho_N=\sup\limits_x |y_N(x)-y(x)|$, то при $\rho_N\to 0$ говорят, что $y_N(x)$ сходится к $y(x)$ равномерно по x; если $\rho_N=\int[y_N(x)-y(x)]^2dx$, то говорят о сходимости $y_N(x)$ к $y(x)$ в среднеквадратическом смысле, и т. д.

Указанные выше требования к сходимости функций могут быть предъявлены как к оценкам функции распределения, так и к оценкам плотности. Однако оказывается, что некоторые оценки плотности, вполне приемлемые для ряда прак-

тических целей, не удовлетворяют ни одному из приведенных выше определений сходимости. Тем не менее, можно говорить об их сходимости в некотором ином смысле.

Будем говорить, что $y_N(x)$ сходится к $y(x)$ в кусочно-интегральном смысле, если для некоторой совокупности интервалов $\{\Delta_k\}$ выполняется условие

$$\int_{\Delta_k} y_N(x)\,dx \to \int_{\Delta_k} y(x)\,dx, \quad k=1, 2,... \qquad (7.1.1)$$

Если (7.1.1) имеет место лишь для конечных Δ_k, будем говорить о слабой кусочно-интегральной сходимости; если (7.1 1) остается справедливым при некотором стремлении интервалов $\Delta_k(N)$ к нулю при $N\to\infty$, то такую кусочно-интегральную сходимость будем называть сильной. Следует отличать случай, когда интервалы $\{\Delta_k(N)\}$ являются случайными и стремятся к нулю в некотором статистическом смысле, от регулярного, неслучайного стремления $\Delta_k(N)$ к нулю, например, как $N^{-\alpha}$, где $\alpha>0$. Как мы увидим впоследствии, при некоторых значениях α сильная регулярная кусочно-интегральная сходимость оказывается эквивалентной равномерной сходимости.

Как мы уже отмечали, различные оценки функции распределения $F_N(x)$ и плотности $p_N(x)$ могут быть получены прямым путем по выборке $x_1, ..., x_N$ на основе того или иного непараметрического факта. Учитывая, однако, что $F(x)$ и $p(x)$ связаны известным образом, можно, получив оценку одной из этих функций, косвенным путем получить оценку другой, соответственно продифференцировав $F_N(x)$ или интегрируя $p_N(x)$. Как правило, полученные таким образом пары оценок будут обладать различными типами сходимости, что должно учитываться при их практическом использовании.

§ 7.2. ЭМПИРИЧЕСКАЯ ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ

Пусть перед нами стоит задача оценивания функции распределения одномерной непрерывной случайной величины X по выборке $x_1, ..., x_N$. Поскольку функция распределения $F(x)$ определяется как вероятность события $(X\leqslant x)$, то оценка этой вероятности будет оценкой для $F(x)$ в точке x. Известно, что оценкой вероятности события является его относительная частота. В данном случае число «благоприятных исходов» равно числу выборочных значений, не превышающих заданную величину x. Используя функцию сравнения $C(t)$ (см. (5 1.1)), можно записать:

$$F_N(x)=\bar{P}r(X\leqslant x)=\frac{m_x}{N}=\frac{1}{N}\sum_{i=1}^{N} C(x-x_i). \qquad (7.2\ 1)$$

Если теперь рассматривать $F_N(x)$ как функцию x, то (7.2.1) является оценкой функции распределения $F(x)$. Этой оценке присвоено наименование эмпирической функции распределения.

Случайная величина $F_N(x)$ распределена согласно биномиальному распределению, поэтому ее среднее значение и дисперсия легко вычисляются:

$$E(F_N(x)) = F(x), \tag{7.2.2}$$

$$D(F_N)) = \frac{1}{N} F(x)(1-F(x)). \tag{7.2.3}$$

Можно вычислить и ковариацию для значений $F_N(x)$ и $F_N(y)$ (Дарлинг [1]):

$$\operatorname{cov}(F_N(x), F_N(y)) = E(F_N(x) \cdot F_N(y)) - F(x) \cdot F(y) =$$

$$= \frac{1}{N} Z(F(x), F(y)), \tag{7.2.4}$$

где

$$Z(s, t) = \begin{cases} s(1-t), & s \leqslant t, \\ t(1-s), & s \geqslant t. \end{cases} \tag{7.2.5}$$

Приведем некоторые следствия для $F_N(x)$ из классических результатов:

Усиленный закон больших чисел:

$F_N(x) \to F(x)$ с вероятностью 1 для всех x.

Закон повторного логарифма:

$$\lim_{N \to \infty} \sup \frac{|F_N(x) - F(x)|}{\sqrt{\frac{2}{N} \log \log N}} = \sqrt{F(x)(1-F(x))}$$

с вероятностью 1 для всех x.

Многомерная центральная предельная теорема:

Совокупность $\{\sqrt{N}(F_N(x_i) - F(x_i))\}$, $i = 1, 2, \ldots, k$ имеет асимптотически ($N \to \infty$, k фиксировано) нормальное k-мерное распределение со средним 0 и ковариациями $Z(F(x_i), F(x_j))$, где $Z(s, t)$ определяется формулой (7.2.5).

Лемма Гливенко-Кантелли (см. Гливенко [1])

$$\sup_{-\infty < x < \infty} |F_N(x) - F(x)| \to 0 \text{ с вероятностью 1.}$$

Все эти результаты доказаны для выборки $(x_1, x_2, \ldots, x_N)$, состоящей из независимых значений. Однако Штейнхаус [1], а затем Вольфовиц [2] показали, что это требование может быть несколько ослаблено.

Таким образом, $F_N(x)$ сходится к $F(x)$ равномерно по x, что говорит о высоком качестве такой оценки неизвестного распределения.

На практике часто бывает необходимым оценить точность получаемой оценки $F_N(x)$, либо, наоборот, задав необходимую точность, указать объем выборки, при котором гарантируется указанная точность. Определяющим моментом здесь является выбор меры точности, так как различные «расстояния» между $F_N(x)$ и $F(x)$ обладают разными статистическими свойствами. Например, если в в качестве меры точности выбрать величину

$$d_N = \sup_{-\infty < x < \infty} |F_N(x) - F(x)|, \tag{7.2.6}$$

то ее асимптотическое распределение выражается формулой:

$$\lim_{N\to\infty} P_r\,(\sqrt{N}\,d_N \leqslant z) = \sum_{-\infty}^{\infty} (-1)^i e^{-2i^2z^2},\ z > 0,$$

с помощью которой можно найти необходимые характеристики точности (например, доверительный интервал, дисперсию и т. п.). Другие меры различия между $F_N(x)$ и $F(x)$ и соответствующие точные и асимптотические распределения этих мер можно найти в работе Залера [1].

Оценка плотности $p_N(x)$, получаемая из $F_N(x)$ путем дифференцирования, аналитически может быть представлена с помощью δ-функций:

$$p_N(x) = \frac{d}{dx} F_N(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i). \tag{7.2.7}$$

Оценка (7.2.7) не обладает свойством равномерной сходимости, но она сходится к $p(x)$ в кусочно-интегральном смысле. В силу своего специфического вида оценка (7.2.7) оказывается очень удобной в тех случаях, когда нужно получить не столько оценку самой функции $p(x)$, сколько оценку среднего той или иной функции случайной величины X. Простейшим примером является оценка математического ожидания:

$$\tilde{E}(X) = \int x p_N(x)\,dx = \frac{1}{N} \sum_i \int x\,\delta(x - x_i)\,dx =$$

$$= \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{7.2.8}$$

На рис. 7.2.1 представлено, как выглядят $F_N(x)$ и $p_N(x)$ графически.

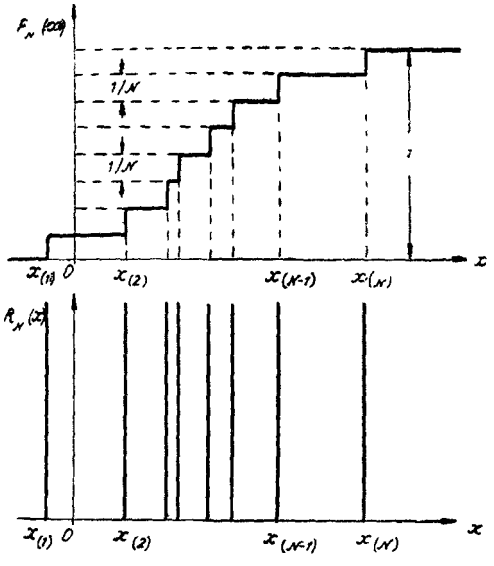

Рис. 7.2.1

§ 7.3. ДИСКРЕТНАЯ (КВАНТИЛЬНАЯ) ОЦЕНКА ФУНКЦИИ РАСПРЕДЕЛЕНИЯ

Для построения эмпирической функции распределения используется непараметрический факт сходимости относительной частоты к неизвестной вероятности. С помощью любого другого непараметрического факта, связанного со значением функции $F(x)$, можно было бы получить другую оценку $F_N(x)$. Такую возможность предоставляет, в частности, статистическая связь между выборочным значением и его рангом, обсужденная в § 5. Одно из проявлений этой связи выражается в том, что

$$E(F(x_i) \mid R_i) = \frac{R_i}{N+1}. \tag{7.3.1}$$

Так как, кроме того, дисперсия величины $F(x_i) - \frac{R_i}{N+1}$ убывает с ростом N, функцию $F(x)$ можно оценить в выборочных точках $\{x_i\}$ величиной $R_i/(N+1)$:

$$F_N(x_i) = \frac{R_i}{N+1}, \quad i=1, 2,..., N. \tag{7.3.2}$$

Вводимую таким образом оценку можно назвать квантильной, поскольку (7.3.2) фактически сводится к исполь-

зованию R-й порядковой статистики в качестве оценки квантиля уровня $R/(N+1)$.

Квантильная оценка функции распределения является дискретной, так как она определена только в точках, соответствующих выборочным значениям $\{x_i\}$. Можно было бы поставить вопрос о «натягивании» некоторой непрерывной функции на квантильную оценку $\{F_N(x_i)\}$, например, о построении полинома, график которого проходит через все точки $\{F_N(x_i)\}$. Однако мы не будем делать такого привнесения произвольных операций в алгоритм вычисления оценки $F_N(x)$; построение оценки должно быть естественным, т. е. основываться на статистических фактах, связанных с природой задачи. В дальнейшем (см. § 7.5) мы увидим, что интерполяция квантильной оценки действительно может быть произведена определенным естественным образом.

§ 7.4. ГИСТОГРАММА

В тех случаях, когда желательно по данным эксперимента построить оценку плотности вероятностей, экспериментаторы чаще всего прибегают к построению гистограммы. Процедура ее построения проста и состоит из следующих шагов.

1. Область возможных значений измеряемой величины X разбивается пороговыми значениями $\{x_k\}$ на прилегающие (не перекрывающиеся и обычно одинаковые) интервалы... $(x_{-(k-1)}, x_{-k}]$, ..., $(x_{-1}, x_0]$, $(x_0, x_1]$, ..., $(x_k, x_{k+1}]$, ...

2. Определяется, сколько выборочных значений из общего числа N оказалось в каждом интервале группировки (обозначим полученные числа через $\{m_k\}$).

3. Над каждым из интервалов строится вертикальный прямоугольник площадью m_k/N. Полученная совокупность прямоугольников и называется г и с т о г р а м м о й.

Основанием для использования гистограммы $p_N(x)$ в качестве оценки неизвестной плотности вероятностей $p(x)$ является кусочно-интегральная сходимость $p_N(x)$ к $p(x)$, которая следует из того, что относительная частота m_k/N события $(X \in (x_k, x_{k+1}])$ сходится к его вероятности p_k:

$$\int_{x_k}^{x_{k+1}} p_N(x)\,dx = \frac{m_k}{N} \to p_k = \int_{x_k}^{x_{k+1}} p(x)\,dx.$$

Такой сходимости в ряде практических случаев оказывается достаточно; однако всегда желательно по возможности улучшить оценку при заданных ограничениях. Несколько факторов влияет на качество гистограммы: объем выборки N,

величина интервалов группировки, положение порогов $\{x^q\}$ при заданной величине интервалов. Все осложняется еще и тем, что степень влияния этих факторов зависит от неизвестного экспериментатору до опыта истинного распределения вероятностей $p(x)$. Поэтому на практике гистограммы строят с некоторым учетом свойств полученной выборки (например, величина интервала группировки выбирается так, чтобы не сгладить существенные особенности распределения; объем выборки связывают с тем, чтобы в ячейке с наименьшим числом измерений их насчитывалось не менее пяти; размещение интервалов связывают с положением наименьшего и наибольшего выборочных значений, и т. п., см., например, А. Хальд [1]). Отметим также, что на качество гистограммы влияет и точность измерений.

Теоретическая задача оптимизации гистограммы может быть сформулирована в нескольких вариантах, однако ее решение связано с трудностями, так что конкретных результатов можно добиться лишь при некоторых частных предположениях. Рассмотрим, например, как влияет точность измерений на выбор величины интервала группировки — в достаточно простом (чтобы довести выкладки до конца), но не слишком тривиальном случае (Ф. П. Тарасенко [5]).

Пусть интервал $[0, L]$ возможных значений измеряемой величины X разбивается на n одинаковых промежутков $\Delta(n\Delta=L)$, которые и служат интервалами группировки для гистограммы. Чтобы исключить влияние конечного объема выборки N, предположим, что $N\to\infty$. Если бы при этом точность измерений была бесконечной, то, неограниченно уменьшая Δ с некоторой скоростью, зависящей от N, мы получали бы все улучшающуюся оценку неизвестной плотности $p(x)$. Однако в результате влияния погрешности измерений при достаточно малых Δ выборочные значения начнут часто попадать не в те интервалы, которым принадлежат истинные значения интересующей нас случайной величины. В результате уменьшение Δ после некоторого Δ_0 приведет к ухудшению получаемой гистограммы. Естественно возникает вопрос об оптимальной величине Δ_0. Оптимальность прежде всего предполагает использование некоторого критерия оптимальности. Возьмем в качестве такого критерия максимум количества информации, которое приносит каждое занесение выборочного значения в соответствующую ячейку гистограммы.

Необходима дальнейшая конкретизация задачи. Во-первых, предположим, что точность измерений характеризуется равномерным распределением истинного значения X около измеренного значения x_i:

$$p(x/x_i)=\begin{cases}\dfrac{1}{2\delta}, & x\in[x_i-\delta,\ x_i+\delta],\\ 0, & x\notin[x_i-\delta,\ x_i+\delta].\end{cases}$$

Таким образом, δ является параметром точности; экспериментаторы часто обозначают это соотношением «$x=x_i\pm\delta$».

Во-вторых, чтобы можно было пренебречь краевыми эффектами, предположим, что точность измерений достаточно велика ($\delta\ll L$) и что ячейки гистограммы достаточно многочисленны ($n=\dfrac{L}{\Delta}\gg 1$).

Априорная неопределенность принадлежности X к какому-то из интервалов исчисляется энтропией $H=-\sum\limits_k p(k)\ln p(k)$, где $p(k)=\int p(x)\,dx$, $p(x)$ — априорное распределение X.

Апостериорная энтропия отлична от нуля, поскольку после занесения измеренного значения x_i в конкретную ячейку остается ненулевой вероятность того, что истинное значение X на самом деле относится к одному из соседних интервалов группировки.

При сделанных выше предположениях (дополнив их слабой изменчивостью $p(x)$ в пределах одного интервала группировки) можно показать, что при $\Delta\geqslant\delta$

$$p_k=Pr(X\in\Delta_k/x_i\in\Delta_k)=1-\frac{\delta}{2\Delta},$$

$$p_{k-1}=Pr(X\in\Delta_{k-1}/x_i\in\Delta_k)=\frac{\delta}{4\Delta},$$

$$p_{k+1}=Pr(X\in\Delta_{k+1}/x_i\in\Delta_k)=\frac{\delta}{4\delta}, \qquad (7.4.1)$$

а при $\dfrac{\delta}{2}\leqslant\Delta\leqslant\delta$

$$p_k=\frac{\Delta}{2\delta},\quad p_{k-1}=p_{k+1}=\frac{1}{2}\left(1-\frac{\Delta}{2\delta}\right). \qquad (7.4.2)$$

Количество информации о случайной величине X, получаемое в результате занесения x_i в определенный интервал группировки, запишется теперь в следующем виде:

$$I=-\sum p(k)\ln p(k)+p_{k-1}\ln p_{k-1}+p_k\ln p_k+p_{k+1}\ln p_{k+1}. \qquad (7.4.3)$$

Далее проблема становится чисто технической: нужно задаться каким-либо конкретным априорным распределением $p(x)$, записать (7.4.3) в явном виде, используя (7.4.1) и

(7.4.2), и найти значение Δ_0, обеспечивающее наибольшую величину I. Не воспроизводя всех выкладок, заметим, что во всем диапазоне априорных распределений (от равномерного в L до δ-образного) наибольшее значение I достигается при $\Delta=\delta$. Очевидно, это связано с инверсией зависимости вероятностей p_k от отношения δ/Δ (ср. (7.4.1) и (7.4.2)).

§ 7.5. ПОЛИГРАММА

Можно предложить (Тарасенко [3]) еще одну непараметрическую оценку плотности, основой для которой является непараметрический факт статистической эквивалентности выборочных блоков (§ 4.3). Одно из конкретных проявлений этого факта состоит в том, что

$$\Pr\left\{\left|[F(x_{(R+K)})-F(x_{(R)})]-\frac{K}{N+1}\right|>\varepsilon\right\}\underset{N\to\infty}{\to}0, \tag{7.5.1}$$

где $1\leqslant K<N$, $F(x)$ — истинная (и неизвестная) непрерывная функция распределения, $x_{(S)}$ — S-я порядковая статистика выборки x_1, ..., x_N. Формула (7.5.1) следует из неравенства Чебышева и формул (7.6.8) из Уилкса [1]. Так как по определению

$$f(x)=\lim_{\Delta\to 0}\frac{F(x+\Delta)-F(x)}{\Delta}, \tag{7.5.2}$$

то возникает естественное предложение воспользоваться в качестве оценки неизвестной плотности в интервале $\Delta_{RK}=(x_{(R+K)}, x_{(R)}]$ величиной

$$f_N(x)=\frac{E[F(x_{(R+K)})-F(x_{(R)})]}{x_{(R+K)}-x_{(R)}}=\frac{K}{(N+1)(x_{(R+K)}-x_{(R)})}, \tag{7.5.3}$$

имея в виду, что $x_{(R+K)}-x_{(R)}\to 0$ при $N\to\infty$ и $K=o(N)$.

Для удобства обозначим порядковые статистики, ранги которых кратны K, через $\xi_j=x_{(jk)}$. Введя обозначение $m=K/N$ и введя функцию-индикатор интервала с помощью ступенчатой функции $c(t)=\{1: t\geqslant 0;\ 0: t<0\}$, результирующую оценку плотности можно записать как

$$f_N(x)=\frac{1}{m}\sum_{j=1}^{m}\frac{c(x-\xi_j)-c(x-\xi_{j-1})}{\xi_{j-1}-\xi_j}. \tag{7.5.4}$$

Будем называть оценку (7.5.4) полиграммой K-го порядка. Построение полиграммы сводится к упорядочению выборки и построению прямоугольников площади $\frac{K}{N+1}$ на примы-

кающих друг к другу интервалах Δ_{jk}. Данный параграф посвящен исследованию асимптотических свойств полиграммы.

Вычислим плотность распределения величины $z=f_N(x)$ для произвольного заданного значения x, удовлетворяющего условию $f(x)>0$. Пусть совместная плотность для величин $\xi_1, \ldots, \xi_m$ обозначена через $p(\xi_1, \ldots, \xi_k)$, причем $\xi_1, \ldots, \xi_k$ принадлежат симплексу $R^m: -\infty<\xi_1<\ldots<\xi_k<\infty$. Плотность распределения величины z запишется в виде

$$p(z)=\int\limits_{R^m}\delta\left(z-\frac{1}{m}\sum_{j=1}^{m}\cdot\frac{c(x-\xi_j)-c(x-\xi_{j+1})}{\xi_{j+1}-\xi_j}\right)\times$$
$$\times\, p(\xi_1,\ldots,\xi_k)\,d\xi_1,\ldots,d\xi_k.$$

Учитывая, что в сумме, выражающей полиграмму, для заданных x и $\{\xi_j\}$ только один член отличен от нуля, имеем:

$$p(z)=\sum_{j=1}^{m}\int\limits_{x}^{\infty}\int\limits_{-\infty}^{x}\delta\left(z-\frac{1/m}{\xi_{j+1}-\xi_j}\right)p(\xi_j,\xi_{j+1})\,d\xi_j\,d\xi_{j+1}, \quad (7.5.5)$$

где (см. гл. IV)

$$p(\xi_j,\xi_{j+1})=\frac{\Gamma(N+1)}{\Gamma(jK)\,\Gamma(K)\,\Gamma(N-(j+1)\,K+1)}[F(\xi_j)]^{jk-1}\times$$
$$\times[F(\xi_{j+1})-F(\xi_j)]^{k-1}\,[1-F(\xi_{j+1})]^{N-(j+1)K}\,f(\xi_j)\,f(\xi_{j+1}). \quad (7.5.6)$$

Производя в (7.5.5) замену переменных $v=F(\xi_{j+1})$, $u=F(\xi_j)$, имеем

$$p(z)=\int\limits_{F(x)}^{1}\int\limits_{0}^{F(x)}\delta\left(z-\frac{1/m}{F^{-1}(v)-F^{-1}(u)}\right)\times$$
$$\times\sum_{j=1}^{m}\frac{\Gamma(N+1)}{\Gamma(jK)\,\Gamma(K)\,\Gamma(N-(j+1)\,K+1)}u^{jk-1}\,(v-u)^{k-1}$$
$$\times\,(1-v)^{N-(j+1)K}\,du\,dv. \quad (7.5.7)$$

Выражение (7.5.7) является точным, однако его вычисление в конечной форме затруднительно. Поэтому перейдем к рассмотрению асимптотического поведения полиграммы в предположении, что $0<f(x)\leqslant C<\infty$, $|f'(x)|<\infty$, $N\to\infty$. Рассмотрим сначала случай фиксированного K. Вынося за знак суммы не зависящий от j сомножитель $(v-u)^{K-1}/\Gamma(K)$, представляя гамма-функции формулой Стирлинга и воспользовавшись стандартной методикой аппроксимации выражений полученного типа (Коваленко и Филиппова [1], стр. 56—59), рассматриваемую сумму можно привести к следующему виду:

$$\frac{N^{K+1}e^{-N(v-u)}(v-u)^{K-1}}{\Gamma(K)\sqrt{2\pi N(1-v)\,u}}\sum_{j=1}^{m}\exp\left\{-\frac{1}{2N}\left[\frac{(jK-1-Nu)^2}{u}+\right.\right.$$
$$\left.\left.+\frac{(N-(j+1)K-N(1-v))^2}{1-v}\right\}(1+0(N^{-\frac{1}{2}})).\qquad(7.5.8)\right.$$

Введя переменную $t_j=\frac{j+1}{m}$ и перейдя от суммирования по j к интегрированию по t, получим (с точностью до величин порядка $0(N^{-\frac{1}{2}})$)

$$p(z)=\int_{F(x)}^{1}\int_{0}^{F(x)}\delta\left(z-\frac{1/m}{F^{-1}(v)-F^{-1}(u)}\right)\frac{N^{K+2}(v-u)^{K-1}}{K\Gamma(K)\sqrt{1-(v-u)}}\times$$
$$\times\exp\left\{-N(v-u)-\frac{N}{2}\cdot\frac{u(v-u)(1-v)}{1-(v-u)}\left[(v-u)-2\,\frac{K-1}{N}\right]\right\}\times$$
$$\times\left\{\Phi\left[\sqrt{\frac{N}{u(1-v)}}\left(\sqrt{1-(v-u)}-\frac{u}{\sqrt{1-(v-u)}}\right)\right]-\right.$$
$$-\Phi\left[\sqrt{\frac{N}{u(1-v)}}\left(\frac{1}{m}\sqrt{1-(v-u)}-\right.\right.$$
$$\left.\left.\left.-\frac{u}{\sqrt{1-(v-u)}}\right]\right\}du\,dv.\qquad(7.5.9)\right.$$

Далее техника вычислений такова. Представим (7.5.9) как

$$\int_{F}^{1}\int_{0}^{F}\delta\left(z-\frac{1/m}{F^{-1}(v)-F^{-1}(u)}\right)\psi(u,v)\,du\,dv,$$

и введя замену переменных $\eta=F^{-1}(v)$, $\xi=F^{-1}(u)$, получим:

$$\int_{x}^{\infty}\int_{-\infty}^{x}\delta\left(z-\frac{1/m}{\eta-\xi}\right)\psi(F(\xi),F(\eta))\,f(\xi)\,f(\eta)\,d\xi\,d\eta.\quad(7.5.10)$$

Сделав в (7.5.10) еще одну замену $t=\xi$, $y=\frac{1/m}{\eta-\xi}$ и проинтегрировав по y, получим:

$$p(z)=\frac{1}{Kmz^2}\int_{x-1/mz}^{x}\psi\cdot f(t)\,f\left(t+\frac{1}{mz}\right)dt.\qquad(7.5.11)$$

Учитывая, что $m=(N+1)/K$, и воспользовавшись теоремой о среднем, с точностью до членов порядка $N^{-\frac{1}{2}}$, получим

$$p(z)=\frac{K^k}{\Gamma(K)\,z}\left(\frac{f}{z}\right)^{K+1}e^{-K\frac{f}{z}}.\qquad(7.5.12)$$

Теорема 7.5.1. Пусть по выборке объема N независимых и одинаково распределенных наблюдений над случайной величиной X, подчиняющейся распределению с плотностью $f(x)$, полиграммой K-го порядка оценивается $f(x)$. Тогда, если $f(x)\leqslant C<\infty$, $|f'(x)|<\infty$ для всех x в области $f(x)>0$, полиграмма имеет асимптотическое (по N) распределение (7.5.12).

Следствие. r-й момент распределения равен

$$Ez^r=f^r K^{r-1}\frac{(K-r)!}{(K-1)!}. \tag{7.5.13}$$

Отсюда следует, что полиграмма K-го порядка ($K\geqslant 1$) является асимптотически несмещенной оценкой $f(x)$ (так как $Ez=f$) и имеет конечные моменты лишь при $r\leqslant K$. Независимость моментов от объема выборки N означает, что при конечных K полиграмма не является состоятельной оценкой $f(x)$.

Рассмотрим теперь случай, когда порядок полиграммы возрастает вместе с объемом выборки, $K=N^{\alpha}$, $0<\alpha<1$. Асимптотическое поведение $p(z)$ при этом, естественно, изменится. Вновь используя формулу Стирлинга, сумму в формуле (7.5.7) можно представить как

$$\frac{N}{2\pi e^2}\sum_{j=1}^{m}\frac{\left[\left(\frac{u}{\alpha_j}\right)^{\alpha_j}\left(\frac{v-u}{\beta}\right)^{\beta}\left(\frac{1-v}{\gamma_j}\right)^{\gamma_j}\right]^N}{\sqrt{\alpha_j\beta\gamma_j}}\left(1+0\left(\frac{1}{K}\right)\right), \tag{7.5.14}$$

где

$$\alpha_j=\frac{jK-1}{N},\quad \beta=\frac{K-1}{N},\quad \gamma_j=\frac{N-(j+1)K}{N}.$$

Снова используя стандартную методику асимптотической аппроксимации таких выражений и учитывая, что $\alpha_j+\beta+\gamma_j=1-\frac{2}{N}$, (7.5.14) можно записать в виде

$$\frac{N}{2\pi\sqrt{u(v-u)(1-v)}}e^{-\frac{N}{2}\frac{(\beta-(v-u))^2}{v-u}}\sum_{j=1}^{m}\exp\left\{-\frac{N}{2}\left[\frac{(j\beta-u)^2}{u}+\right.\right.$$
$$\left.\left.+\frac{(v-(j+1)\beta)^2}{1-v}\right]\right\}\left(1+0\left(\frac{1}{K}\right)\right). \tag{7.5.15}$$

Переходя к суммированию по переменной $t_j=j\frac{K}{N}$ и заменяя суммирование интегрированием по t, имеем

$$p(z)=\int_F^1\int_0^F\delta\left(z-\frac{1/m}{F^{-1}(v)-F^{-1}(u)}\right)\frac{m\sqrt{N}}{\sqrt{2\pi(v-u)(1-(v-u))}}\times$$

$$\times \exp\left\{-\frac{N}{2}\left[\frac{(\beta-(v-u))^2}{v-u}+\frac{(v-u)^2}{1-(v-u)}\right]\right\}\times$$

$$\times\left\{\Phi\left[\sqrt{N}\left(\sqrt{\frac{1-(v-u)}{u(1-v)}}-\frac{1}{1-v}\sqrt{\frac{u(1-v)}{1-(v-u)}}\right)\right]-\right.$$

$$\left.-\Phi\left[\sqrt{N}\left(\frac{1}{m}\sqrt{\frac{1-(v-u)}{u(1-v)}}-\frac{1}{1-v}\sqrt{\frac{u(1-v)}{1-(v-u)}}\right)\right]\right\}dudv. \tag{7.5.16}$$

Проделав операции, аналогичные переходу от (7.5.9) к (7.5.11), получим

$$p(z)=\frac{\left(\frac{f}{z}\right)^{5/2}}{f\sqrt{\frac{2\pi}{K}}}\exp\left\{-\frac{1}{2}\left[\frac{\left(1-\frac{f}{z}\right)^2}{\frac{f}{Kz}}+\frac{K}{m}\frac{\left(\frac{f}{z}\right)^2}{1-\frac{f}{mz}}\right]\right\}\times$$

$$\times(1+0(K^{-\frac{1}{2}})). \tag{7.5.17}$$

Легко видеть, что при $K=0(m)$ (т. е. $0<\alpha<1/2$) случайная величина $t_N=\sqrt{N^{\alpha}}\left(\sqrt{\frac{z}{f}}-\sqrt{\frac{f}{z}}\right)$ имеет асимптотически нормальное стандартное распределение

Таким образом, справедлива следующая

Теорема 7.5.2. Если неизвестная плотность $f(x)>0$, ограниченная вместе со своей первой производной, оценивается полиграммой $f_N(x)$ K-го порядка, $K=N^{\alpha}$, $0<\alpha<1/2$, то $f_N(x)$ является состоятельной оценкой $f(x)$, а случайная величина $t_N=\sqrt{N^{\alpha}}(\sqrt{f_N/f}-\sqrt{f/f_N})$ распределена асимптотически нормально с нулевым средним и единичной дисперсией.

§ 7.6. ОЦЕНКА НЕИЗВЕСТНОЙ ПЛОТНОСТИ С ПОМОЩЬЮ РАЗЛОЖЕНИЯ ПО ВЕСОВЫМ ФУНКЦИЯМ

Было бы весьма желательно иметь такую оценку неизвестной плотности, которая обладала бы рядом хороших качеств, например, несмещенностью и быстрой сходимостью во всех точках x. Нереально, однако, требовать от природы больше того, что она может дать. Например, можно показать (М. Розенблат [1]), что при достаточно общей постановке задачи не существует несмещенных оценок неизвестной плотности $p(x)$.

Допустим обратное, т. е. что существует такая статистика $S(x; x_1, ..., x_N)=p_N(x)$, что

$$ES(x; X_1, ..., X_N)\equiv p(x). \tag{7.6.1}$$

Потребуем, чтобы $S(x; x_1, ..., x_N)$ было положительной и симметричной по переменным $x_1, ..., x_N$ функцией. Тогда

$$\int_a^b S(x; x_1, ..., x_N)\, dx$$

является симметричной по $x_1, ..., x_N$ и несмещенной оценкой величины

$$F(b)-F(a)=\int_a^b p(x)\, dx.$$

Действительно, в силу (7.6.1)

$$E\int_a^b S(x; X_1, ..., X_N)\, dx = \int_a^b ES(x; X_1, ..., X_N)\, dx=$$

$$=\int_a^b p(x)\, dx=F(b)-F(a).$$

Единственной несмещенной и симметричной оценкой вероятности является относительная частота, т. е. $F_N(b)-F_N(a)$, где $F_N(x)$ — эмпирическая функция распределения (7.2.1). Таким образом,

$$F_N(b)-F_N(a)=\int_a^b S(x; x_1, ..., x_N)\, dx \qquad (7.6.2)$$

для всех a и b и почти всех $x_1, ..., x_N$. Отсюда следует, что $F_N(x)$ должно быть абсолютно непрерывным * по x, что не соответствует действительности и тем самым опровергает предположение (7.6.1) о возможной несмещенности $p_N(x)$.

Однако существует класс оценок $p_N(x)$ плотности $p(x)$, удовлетворяющих условию равномерной по x сходимости к $p(x)$ при $N\to\infty$. Впервые такие оценки были предложены в 1956 году Розенблатом [1]; в последующие годы они подвергались интенсивному исследованию ряда авторов. Дадим краткое изложение полученных в этом направлении результатов.

Первоначальная идея построения равномерно сходящейся оценки функции $p(x)$ исходит из определения плотности:

$$p(x)=\frac{d}{dx}F(x)=\lim_{h\to 0}\frac{F(x+h)-F(x-h)}{2h}. \qquad (7.6.3)$$

---

* Функция f, заданная на некотором отрезке $[a, b]$, называется абсолютно непрерывной на этом отрезке, если для любого $\varepsilon>0$ найдется такое $\delta>0$, что какова бы ни была конечная система попарно непересекающихся интервалов (a_k, b_k), $k=1, 2, ..., n$, принадлежащих $[a, b]$, и такая, что при $\Sigma(b_k-a_k)<\delta$, выполняется неравенство $\Sigma|f(b_k)-f(a_k)|<\varepsilon$ (А. Н. Колмогоров и С. В. Фомин [1]).

Соотношение (7.6.3) подсказывает, что, взяв достаточно малый интервал h при достаточно большом N, можно надеяться на получение достаточно хорошей оценки плотности

$$p_N(x)=\frac{F_N(x+h)-F_N(x-h)}{2h}; \tag{7.6.4}$$

если же $N\to\infty$, то h можно устремить к нулю с некоторой скоростью, $h(N)\to 0$ (так, чтобы число выборочных значений в $(x-h,\ x+h)$ неограниченно возрастало), тем самым сколь угодно приближаясь к (7.6.3).

Необходимо конкретизировать зависимость $h(N)$, обеспечивающую ожидаемую сходимость. Для удобства анализа придадим (7.6.4) несколько иную форму. Воспользовавшись явной записью $F_N(x)$ (7.2.1), имеем

$$p_N(x)=\frac{1}{2Nh}\left[\sum_{i=1}^{N} C(x+h-x_i)-\sum_{i=1}^{N} C(x-h-x_i)\right]=$$

$$=\frac{1}{2h}\left\{\int_{-\infty}^{x+h} dF_N(t)-\int_{-\infty}^{x-h} dF_N(t)\right\}=\frac{1}{2h}\int_{x-h}^{x+h} dF_N(t)=$$

$$=\frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{x-t}{h}\right) dF_N(t), \tag{7.6.5}$$

где

$$K(y)=\begin{cases}\dfrac{1}{2}, & |y|\leqslant 1,\\ 0, & |y|>1.\end{cases} \tag{7.6.6}$$

Так мы приходим к оценке вида

$$p_N(x)=\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right). \tag{7.6.7}$$

Теперь обобщение напрашивается само собой: можно рассматривать не только $K(y)$ вида (7.6.6), но и любую функцию $K(y)$ из некоторого класса функций. Если (7.6.6) обеспечивает вычисление относительной частоты попадания X в заданный интервал, то другие функции $K(y)$ дадут как бы «взвешенную относительную частоту», что и оправдывает введение названий «весовая функция» для $K(y)$ и «оценка разложением по весовым функциям» для $p_N(x)$ (7.6.7).

Некоторые ограничения на функции $K(y)$ можно наложить сразу, по очевидному соображению, что $K(y)$ должна обладать всеми свойствами плотности вероятностей:

$$K(y)\geqslant 0, \tag{7.6.8}$$

$$\int_{-\infty}^{\infty} K(y)\,dy=1. \tag{7.6.9}$$

Другие ограничения на $K(y)$ будут вытекать из требований определенной сходимости $p_N(x)$ к $p(x)$. (При этом нам придется наложить также некоторые ограничения и на класс оцениваемых функций $p(x)$).

Потребуем сначала, чтобы оценка была асимптотически несмещенной:

$$\lim_{N\to\infty} Ep_N(x)=p(x). \tag{7.6.10}$$

Так как мы считаем, что все X_t одинаково распределены, то

$$\begin{aligned} Ep_N(x) &= \frac{1}{N}\sum_{t=1}^{N}\left\{E\left[\frac{1}{h}K\left(\frac{x-X_t}{h}\right)\right]\right\}= \\ &= E\left[\frac{1}{h}K\left(\frac{x-X_t}{h}\right)\right]= \\ &= \int_{-\infty}^{\infty}\frac{1}{h(N)}K\left(\frac{x-t}{h(N)}\right)p(t)\,dt. \end{aligned} \tag{7.6.11}$$

Условия, при которых (7.6.11) подчиняется требованию (7.6.10), даются следующей теоремой:

Теорема 7.6.1 (Бохнер [1], Парзен [1]). Если $K(y)$ удовлетворяет условиям (7.6.8) и (7.6.9), а также

$$\sup_{-\infty<y<\infty} K(y)<\infty, \tag{7.6.12}$$

$$\lim_{|y|\to\infty} yK(y)=0; \tag{7.6.13}$$

если $h(N)$ — последовательность положительных величин такая, что

$$\lim_{N\to\infty} h(N)=0, \tag{7.6.14}$$

то в любой точке непрерывности функции $p(x)$ имеет место (7.6.10).

Доказательство. Введем обозначение $Ep_N(x)=g_N(x)$. Разность $g_N(x)-p(x)$ может быть представлена в виде:

$$g_N(x)-p(x)=\int_{-\infty}^{\infty}[p(x-t)-p(x)]\frac{1}{h(N)}K\left(\frac{t}{h(N)}\right)dt. \tag{7.6.15}$$

Пусть $\delta>0$; разобьем область интегрирования на две области, $|t|\leqslant\delta$ и $|t|>\delta$. Тогда имеем:

$$|g_N(x)-p(x)|\leqslant\max_{|t|<\delta}|p(x-t)-p(x)|\cdot\int_{|t|<\delta}\frac{1}{h(N)}K\left(\frac{t}{h(N)}\right)dt+$$

$$+\int_{|t|\geqslant\delta}\frac{|p(x-t)|}{t}\cdot\frac{t}{h(N)}\cdot K\left(\frac{t}{h(N)}\right)dt+p(x)\int_{|t|<\delta}\frac{1}{h(N)}K\left(\frac{t}{h(N)}\right)dt.$$

Так как

$$\int_{|t|<\delta}\frac{1}{h(N)}K\left(\frac{t}{h(N)}\right)dt=\int_{|z|<\frac{\delta}{h(N)}}K(z)\,dz\leqslant\int_{-\infty}^{\infty}K(z)\,dz=1,$$

$$\int_{|t|\geqslant\delta}\frac{1}{h(N)}K\left(\frac{t}{h(N)}\right)dt=\int_{|z|>\frac{\delta}{h(N)}}K(z)\,dz\to 0\text{ при }N\to\infty,$$

$$|g_N(x)-p(x)|\leqslant\max_{|t|<\delta}|p(x-t)-p(x)|+\frac{1}{\sigma}\sup_{|t|>\frac{\delta}{h(N)}}(tK(t))+$$

$$+p(x)\int_{|z|>\frac{\delta}{N(t)}}K(z)\,dz.$$

Правая часть последнего неравенства стремится к нулю при $N\to\infty$ и $\delta\to 0$, что и доказывает теорему об асимптотической несмещенности оценки $p_N(x)$ во всех точках непрерывности функции $p(x)$.

Если $p(x)$ имеет производные в точке x, то суждения о характере смещения оценки $p_N(x)$ при конечных N можно продвинуть дальше. Соотношение (7.6.15) может быть простой заменой переменных приведено к виду

$$g_N(x)-p(x)=\int[p(x-h(N)\,y)-p(x)]\,K(y)\,dy,$$

что при достаточно малых h может быть представлено рядом:

$$g_N(x)-p(x)=hp'(x)\cdot\int yK(y)\,dy+$$

$$+\frac{1}{2}h^2p''(x)\int y^2K(y)\,dy+O(h^3). \qquad (7.6.16)$$

Из (7.6.16) следуют дополнительные желательные свойства функций $K(y)$: плотность $K(y)$ должна иметь все моменты; желательно также, чтобы $\int K(y)\cdot y\,dy=0$, т. е. чтобы $K(y)$ было симметричной относительно нуля функцией.

Рассмотрим теперь поведение дисперсии оценки $p_N(x)$

$$\mathrm{Var}[p_N(x)]=\frac{1}{N}\mathrm{Var}\left[\frac{1}{h}K\left(\frac{x-X}{h}\right)\right]=$$

$$=\frac{1}{Nh^2}\int_{-\infty}^{\infty}K^2\left(\frac{x-y}{\quad}\right)p(y)\,dy=\frac{1}{Nh}\int_{-\infty}^{\infty}K^2(t)\,p(x-ht)\,dt.$$

Следовательно, при достаточно малых h,

$$\mathrm{Var}[p_N(x)] \sim \frac{p(x)}{Nh(N)} \int_{-\infty}^{\infty} K^2(t)\,dt. \qquad (7.6.17)$$

Появляется еще одно желательное свойство $K(y)$ —

$$\int_{-\infty}^{\infty} K^2(y)\,dy < \infty.$$

Кроме того, требование стремления дисперсии оценки к нулю при $N \to \infty$ приводит к новому ограничению на последовательность $h(N)$, кроме (7.6.14) должно иметь место:

$$\lim_{N\to\infty} Nh(N) = \infty. \qquad (7.6.18)$$

Таким образом, хотя $K\left(\frac{x-x_i}{h(N)}\right)$ при $N \to \infty$ устремляется к δ-функции, но не слишком быстро, так, чтобы число выборочных значений, «взвешиваемых» каждой функцией $K\left(\frac{x-x_i}{h}\right)$, неограниченно возрастало. При этом $p_N(x)$ в каждой точке x оказывается суммой неограниченно возрастающего числа случайных величин $V_i = [1/h(N)] \times K[(x-X_i)/h(N)]$, и, следовательно (см. гл VI), при определенных условиях последовательность $p_N(x)$ может подчиняться центральной предельной теореме, т. е. оценка $p_N(x)$ может быть асимптотически нормальной:

$$\lim_{N\to\infty} P\left[\frac{p_N(x) - Ep_N(x)}{\sigma[p_N(x)]} \leqslant c\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c} e^{\frac{y^2}{2}}\,dy = \Phi(c) \qquad (7.6.19)$$

для любого вещественного c. Парзен [1] показал, что достаточным для этого условием является существование при некотором $\delta > 0$ предела

$$\frac{E|V_N - EV_N|^{2+\delta}}{N^{\frac{\delta}{2}}\,\sigma^{2+\delta}[V_N]} \to 0 \text{ при } N \to \infty. \qquad (7.6.20)$$

Высокие качества оценок с помощью разложения по весовым функциям привлекли к ним внимание многих авторов. Можно выделить три направления исследований: 1) ослабление ограничений на условия сходимости $p(x)$ к $p(x)$; 2) оптимизация разложения (7.6.7); 3) вопросы применения данного класса оценок в различных статистических процедурах.

Э. Надарая [1] показал, что для того, чтобы с вероятностью единица при $N \to \infty$ имело место

$$\sup_{-\infty<x<\infty} |p_N(x)-p(x)| \to 0,$$

необходимо и достаточно, чтобы $K(x)$ была функцией с ограниченным изменением, $p(x)$ — равномерно непрерывной функцией и ряд $\sum_{k=1}^{\infty} \exp[-\gamma k h^2(k)]$ сходился при любом γ (последнее имеет место при $h(N)=N^{-\theta}\, 0<\Theta<\frac{1}{2}$). К. Чанда [1] установил, что если потребовать только асимптотической несмещенности $p_N(x)$ и не налагать на $K(x)$ условия обязательной неотрицательности, то среднеквадратическая ошибка оценки может быть уменьшена по сравнению с (7.6.17). В. Мэрфи [1] показал, что сходимость $p_N(x)$ к $p(x)$ во всех точках непрерывности последней сохраняется и в том случае, когда $F(x)$ имеет счетное число разрывов.

Оптимизация разложения (7.6.7) ведется за счет выбора наилучшей формы функции $K(x)$ и конкретизации зависимости $h(N)$. Конечно, все определяется выбором критерия и условий оптимизации. Например, если минимизировать среднеквадратичную разность между $p_N(x)$ и $p(x)$, усредняя ее по $p(x)$, то оптимальное $K(x)$ будет разным для разных $p(x)$ (см. Уотсон и Лидбеттер [1]). Другой подход, являющийся непараметрическим, состоит в том, чтобы минимизировать дисперсию оценки (7.6.17) за счет минимизации интеграла $\int K^2(t)\,dt$ при некоторых дополнительных ограничениях на $K(t)$. При ограниченности интервала $2h$, в котором $K(t)\neq 0$, и задании $2m$ моментов для $K(t)$ оптимальная форма $K(t)$, очевидно, имеет вид $K_0(t)=\sum_{s=0}^{2m}\lambda_N(s)\cdot t^s$ (см. К. Чанда [1]); при задании только дисперсии для $K(y)$ получаем (см. Епонечников [1]) $K_0(t)=\frac{3}{4\sqrt{5}}-\frac{3}{20\sqrt{5}}\cdot t^2$, при $|t|\leqslant\sqrt{5}$ и $K_0(t)=0$ при $|t|>\sqrt{5}$. Оптимальный же коэффициент $h(N)$ получается в виде $C\cdot N^{-\frac{1}{5}}$.

Вопросы использования оценок типа (7.6.7) в конкретных статистических процедурах будут рассмотрены в некоторых из последующих глав; здесь же отметим только, что, как показали Надарая [1] и Бхаттачарья [1], производные ряда (7.6.7) с вероятностью единица сходятся к производным неизвестного распределения $p(x)$:

$$\sup_{-\infty<x<\infty} |p_N^{(r)}(x)-p^{(r)}(x)| \to 0 \quad \text{при } N\to\infty. \qquad (7.6.21)$$

В заключение необходимо отметить весьма важное обобщение теории оценок плотности разложениями по весовым

функциям, сделанное М. Розенблатом [2]. Оказывается, что даже если выборочные значения $x_1, \ldots, x_N$ не являются независимыми, то при определенных ограничениях на характер зависимости асимптотическое поведение оценки $p_N(x)$ (7.6.7) остается тем же, что и при независимых наблюдениях. Требования к типу зависимости носят несколько специфический характер, но, по существу, сводятся к достаточно быстрому ослаблению зависимости при увеличении $|i-k|$, где i и k — номера выборочных значений; таким требованиям удовлетворяет, например, гауссов стационарный марковский процесс и некоторые другие типы марковских процессов.

§ 7.7. ОЦЕНИВАНИЕ ФУНКЦИИ РАСПРЕДЕЛЕНИЯ И ПЛОТНОСТИ ВЕРОЯТНОСТЕЙ РЯДАМИ ФУРЬЕ

Один из напрашивающихся подходов к оценке неизвестного распределения состоит в том, чтобы представить его в виде ряда по системе ортогональных функций; при этом оценивание $F(x)$ или $p(x)$ сводится к оцениванию по выборке $x_1, \ldots, x_N$ коэффициентов соответствующего ряда. В зависимости от того, какая априорная информация используется для построения оценки коэффициентов, данный подход может носить как параметрический, так и непараметрический характер; нас, естественно, будет интересовать последний случай.

Непараметрическое оценивание функции распределения или плотности состоит в том, чтобы использовать для оценки коэффициентов некоторый непараметрический факт. В данном параграфе мы еще раз убедимся, что различное употребление одного и того же факта может дать оценки, обладающие различными свойствами. Мы рассмотрим вариант оценки коэффициентов ряда Фурье, опирающийся на использование эмпирической функции распределения (Кронмал и Тартер [1]).

Пусть $\{\psi_k(x)\}$ — семейство функций, взаимно ортогональных с весом $w(x)$,

$$\int \psi_k(x)\,\psi_j(x)\,w(x)\,dx=\delta_{kj}, \qquad (7.7.1)$$

где δ_{kj} — символ Кронекера

Определим оценку $F_{Nm}(x)$ функции распределения $F(x)$ как конечный ряд

$$F_{Nm}(x)=\sum_{k=0}^{m} \hat{A}_k\,\psi_k(x); \qquad (7.7.2)$$

в качестве коэффициентов $\{\hat{A}_k\}$ этого ряда будем использовать коэффициенты Фурье эмпирической функции распределения $F_N(x)$ (7.2.1):

$$\hat{A}_k = \int F_N(x)\,\psi_k(x)\,w(x)\,dx = \frac{1}{N}\sum_{i=1}^{N}\int C(x-x_i)\,\psi_k(x)\,w(x)\,dx. \qquad (7.7.3)$$

Обсудим свойства введенной таким образом оценки $F_{Nm}(x)$. Поскольку эмпирическая функция распределения является кусочно-гладкой, то почти везде (за исключением точек разрыва)

$$\lim_{m\to\infty} F_{Nm}(x) = F_N(x), \qquad (7.7.4)$$

если $\{\psi_k(x)\}$ является полной ортонормированной системой функций (см. Г. П. Толстов [1]). Далее, если неизвестная функция $F(x)$ является непрерывной, то

$$E\left(\lim_{m\to\infty} F_{Nm}(x)\right) = E(F_N(x)) = F(x). \qquad (7.7.5)$$

Кроме того,

$$E(\hat{A}_k) = A_k, \qquad (7.7.6)$$

где A_k — коэффициенты Фурье для $F(x)$. В самом деле,

$$E(\hat{A}_k) = E\left[\frac{1}{N}\sum_{i=1}^{N}\int C(x-X_i)\,\psi_k(x)\,w(x)dx\right] =$$

$$= E\left[\int C(x-X_i)\,\psi_k(x)\,w(x)\,dx\right] =$$

$$= -\int\left[\int_{-\infty}^{y}\psi_k(x)\,w(x)\,dx\right]dF(y) =$$

$$= -\left[F(y)\int_{-\infty}^{y}\psi_k(x)\,w(x)\,dx\right]_{-\infty}^{\infty} + \int F(y)\,\psi_k(y)\,w(y)\,dy.$$

Последняя строка получена путем интегрирования по частям. Так как $F(-\infty)=0$ и $\int \psi_k\, w dx = 0$, а $\int F\psi_k w dx = A_k$, то (7.7.6) доказано.

Определим взвешенную среднеквадратическую ошибку оценки $F_{Nm}(x)$ как

$$J(F_{Nm}) = E\int [F(x)-F_{Nm}(x)]^2\,w(x)\,dx. \qquad (7.7.7)$$

Применив к (7.7.7) формулу Парсеваля, легко получить, что

$$J(F_{Nm}) = \int F^2(x)\,w(x)dx - \sum_{k=0}^{m}(A_k^2 - \operatorname{Var}\hat{A}_k). \qquad (7.7.8)$$

Если $\{\psi_k(x)\}$ — полная система функций, то $\int F^2 w dx = \sum_0^\infty A_k^2$, и мы имеем

$$J(F_{Nm}) = \sum_{k=0}^{m}\operatorname{Var}\hat{A}_k + \sum_{k=m+1}^{\infty}A_k^2. \qquad (7.7.9)$$

Учитывая (7.7.5) и (7.7.9), можно показать, что разница между взвешенными среднеквадратическими ошибками $F_{Nm}(x)$ и $F_N(x)$ равна

$$J(F_{Nm}) - J(F_N) = \sum_{k=m+1}^{\infty} (A_k^2 - \operatorname{Var} \hat{A}_k). \qquad (7.7.10)$$

Интересно, что можно выбрать число m членов ряда (7.7.2) так, что (7.7.10) будет отрицательным, т. е. взвешенная среднеквадратическая ошибка F_{Nm} будет меньшей, чем у эмпирической функции распределения F_N.

Перейдем теперь к вопросам оценивания плотности $p(x)$. С одной стороны, мы можем предложить в качестве оценки $p(x)$ конечный ряд такого же типа, что и ряд для $F(x)$:

$$p_{Nm}(x) = \sum_{k=0}^{m} \hat{a}_k \psi_k(x), \qquad (7.7.11)$$

и рассматривать ее как самостоятельную оценку. Свойства этой оценки определяются выбором оценок для $\{\hat{a}_k\}$. Например, если потребовать, чтобы

$$E(\hat{a}_k) = a_k, \qquad (7.7.12)$$

где $\{a_k\}$ — коэффициенты Фурье для истинной (оцениваемой) функции $p(x)$, то оценка (7.7.11) окажется смещенной, так как

$$E(p(x) - p_{Nm}(x)) = \sum_{k=m+1}^{\infty} a_k \psi_k(x), \qquad (7.7.13)$$

что равно нулю лишь при $a_k = 0$ для всех $k > m$. Правда, есть основания полагать, что смещенность является свойством, общим для почти всех оценок плотности (см. Розенблатт [1]). Как и для F_{Nm}, можно вычислить взвешенную среднеквадратическую ошибку p_{Nm}, которая оказывается равной

$$\begin{aligned} J(p_{Nm}) &= E \int [p(x) - p_{Nm}(x)]^2 w(x)\, dx = \\ &= \int p^2(x) w(x)\, dx - \sum_{k=0}^{m} (a_k^2 - \operatorname{Var} \hat{a}_k) = \\ &= \sum_{k=0}^{m} \operatorname{Var}(\hat{a}_k) + \sum_{k=m+1}^{\infty} a_k^2. \end{aligned} \qquad (7.7.14)$$

С другой стороны, можно рассматривать не только самостоятельную оценку (7.7.11), но и оценку $p_{Nm}^*(x)$, являющуюся производной от $F_{Nm}(x)$:

$$p_{Nm}^*(x) = \frac{d}{dx} F_{Nm}(x) = \sum_{k=0}^{m} \hat{A}_k \frac{d}{dx} \psi_k(x). \qquad (7.7.15)$$

Было бы желательно, чтобы (7.7.15) тоже было рядом по ортогональным функциям. Это накладывает определенные ограничения на $\{\psi_k\}$, которые, однако, в ряде случаев можно удовлетворить. Например, этим ограничениям подчиняются такие системы ортогональных функций, как $\{\cos k\pi x\}$, $\{\sin k\pi x\}$, $\{\cos k\pi x, \sin k\pi x\}$.

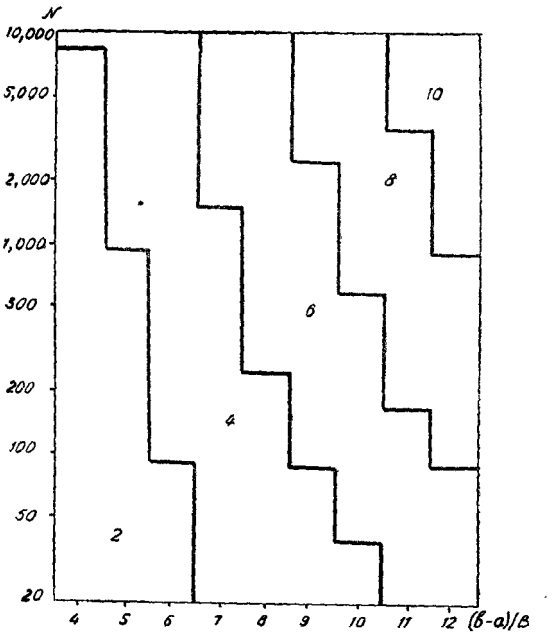

Рис. 7.7.1

При практическом построении оценок $F_{Nm}(x)$ и $p_{Nm}(x)$ возникает ряд дополнительных вопросов. Исследования показывают, например, что для неограниченного убывания взвешенной среднеквадратичной ошибки $p_{Nm}(x)$ достаточно, чтобы число членов в (7.7.11) было порядка $\sqrt{N}$. Особую проблему создает возможность отрицательных значений сумм (7.7.2) и (7.7.11). Можно выбрать специальную систему неотрицательных функций $\{\psi_k\}$, однако это приводит к серьезным практическим усложнениям, а с другой стороны, это оказывается ненужным, так как при объемах выборки уже порядка нескольких десятков появление отрицательных значений у $p_{Nm}(x)$ практически маловероятно (см. Кронмал и Тартер [1]). Далее отметим практическую трудность, типичную для всех аппроксимаций функций конечными рядами: качество оценок F_{Nm} и p зависит от четырех связанных

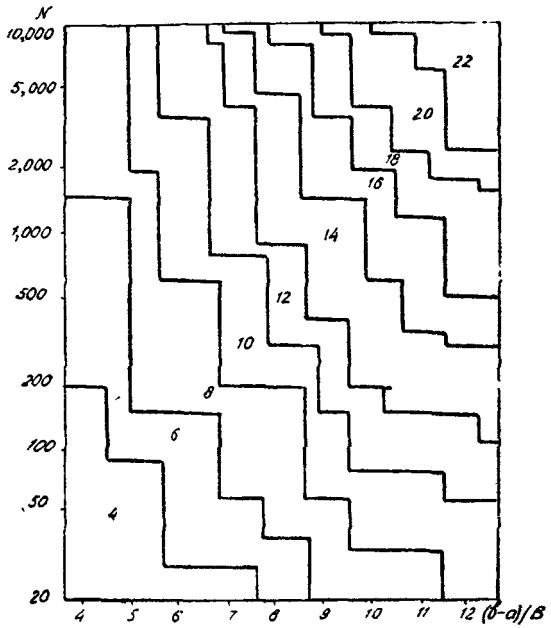

Рис. 7.7 2

между собой факторов: выбора системы функций $\{\psi_k(x)\}$, выбора числа членов ряда m, характера истинной оцениваемой функции F (или p) и относительной величины интервала, в котором осуществляется оценивание. Некоторое представление о связи этих факторов дают два графика из работы Кронмала и Тартера [1]: на них приведены области, соответствующие оптимальному числу m членов ряда для $p_{Nm}(x)$ при заданных объеме выборки N и величине интервала $(b-a)$, измеряемого в единицах $B=(x_{0,75}-x_{0,25})$, где x_p — квантиль уровня p. График 7.7.1 относится к оцениванию плотности нормального распределения, график 7.7.2 — распределения Коши; в обоих случаях разложение производится по синусам и косинусам периода $(b-a)/B$.

§ 7.8. ОБ ОЦЕНИВАНИИ МНОГОМЕРНЫХ ПЛОТНОСТЕЙ И ФУНКЦИЙ РАСПРЕДЕЛЕНИЯ

Во многих практически важных случаях необходимо производить оценивание многомерных распределений вероятностей. Учитывая направленность данной книги, мы полностью опустим рассмотрение параметрической постановки этой за-

дачи, хотя и в этом случае возникает ряд непростых проблем. Нас будут интересовать вопросы построения непараметрических оценок p-мерных плотности или функции распределения по p-мерной выборке объема N: $z_1, \ldots, z_i, \ldots, z_N$, $z_i \equiv \{(x_{i1}, x_{i2}, \ldots, x_{ip})\}$.

Каждый из изложенных в предыдущих параграфах этой главы методов в принципе может быть обобщен на многомерный случай, и такое обобщение действительно проведено рядом авторов (см. например, Мэрфи [2]). Само введение многомерных оценок распределений так, чтобы одномерные оценки были их частным случаем, обычно не составляет труда; сложность состоит в том, чтобы учесть специфически многомерные особенности. Таких особенностей несколько. Во-первых, хотя отдельные векторы z_i выборочных значений могут быть независимыми между собой, это не исключает возможной статистической зависимости между компонентами x_{i1}, x_{i2}, ..., x_{ip} каждого из выборочных векторов. Поэтому, например, оценка многомерной функции распределения в виде произведения одномерных эмпирических функций распределения является состоятельной лишь при независимости компонент x_{i1}, ..., x_{ip} выборочного вектора; при наличии зависимости p-мерная эмпирическая функция распределения должна определяться иначе (см. § 8.6). Во-вторых, условия, при которых обеспечивается сходимость того или иного типа для p-мерной оценки, естественно, становятся более сложными, чем для одномерной, и требуют специального рассмотрения. Но это уже забота теоретиков; практикам остается лишь сопоставлять установленные теоретически ограничения на применимость той или иной оценки с условиями получения данной выборки и тем самым решать, можно ли использовать некоторую оценку в данном конкретном случае. Что доставляет практикам гораздо больше хлопот, так это третья особенность многомерных оценок — громоздкость и трудоемкость вычислений, необходимых для получения таких оценок. Количество вычислительных операций и объем информации, которую необходимо хранить в течение выполнения этих операций, в многомерном случае резко возрастают; этот рост так силен, что даже при использовании современных электронных вычислительных машин практическое получение многомерных оценок, являющихся прямым обобщением одномерных, в принципе возможное, становится фактически невыполнимым.

Много усилий тратят сейчас исследователи, чтобы если не преодолеть, то обойти эту трудность. Три идеи представляются сейчас перспективными в этом отношении.

1. Изыскание новых оценок многомерных распределений,

которые требуют заведомо меньшего объема вычислений, чем известные ранее. Надежды здесь основываются на поисках новых непараметрических фактов или новых методов использования уже известных непараметрических фактов.

2. Разбиение выборки на части, для каждой из которых вычисления обозримы, и последующее объединение подвыборочных оценок, без перенесения «внутренних», исходных и промежуточных данных из одной подвыборки в другую

3. Построение итеративных оценок, которые, независимо от объема выборки, требуют поочередной обработки многомерных выборочных значений, когда при каждой итерации объем необходимых вычислений остается одним и тем же и когда каждая итерация приводит к (стохастическому) улучшению оценки, полученной на предыдущем шаге.

Не претендуя на полноту охвата имеющихся к данному моменту результатов, проиллюстрируем характер поисков в каждом из указанных направлений на конкретных примерах.

В плане реализации первой идеи — поиска простых многомерных оценок — интересной является работа Лофтсгардена и Квезенберри [1]. Многомерная плотность в точке $z=(x_1, \ldots, x_N)$ может быть определена как

$$f(z)=\lim_{r\to 0}[Pr(Z\subset S_r)/V_r], \tag{7.8 1}$$

где S_r — гиперсфера радиуса r с центром в точке z, V_r объем этой сферы. Успех Лофтсгардена и Квезенберри основан на удачном выборе радиуса гиперсферы, $r=r_{k(N)}$, который, с одной стороны, автоматически устремляется к нулю при $N\to\infty$, а с другой — позволяет получить весьма простую оценку вероятности $Pr(Z\subset S_r)$.

Пусть $\{k(N)\}$ — последовательность положительных целых чисел, такая, что

$$\lim_{N\to\infty} k(N)=\infty \tag{7.8.2}$$

и

$$\lim_{N\to\infty}\frac{k(N)}{N+1}=0. \tag{7.8.3}$$

При заданных N, $k(N)$ и выборке $z_1, \ldots, z_N$, радиус $r_{k(N)}$ определяется как расстояние от точки z, в которой оценивается плотность, до $k(N)$-й ближайшей выборочной точки. Объем гиперсферы, определенной таким радиусом, равен

$$V_{r(k)}=\frac{2\pi^{\frac{p}{2}}\cdot r^p_{k(N)}}{p\Gamma\left(\frac{p}{2}\right)}, \tag{7.8.4}$$

а вероятность $Pr(Z \subset S_r)$ есть, очевидно, сумма $k(N)$ покрытий при выборке гиперсфер в качестве упорядочивающих функций. Поэтому в качестве оценки $Pr(Z \subset S_r)$ можно взять величину $k(N)/(N+1)$ (см. гл. IV), которая в силу (7.8.3) стремится к нулю при $N \to \infty$. Но при этом и $r_{k(N)}$, а с ним и V_r сходится к нулю. Это и обеспечивает сходимость оценки

$$p_N(x_1, \ldots, x_p) = \frac{p \cdot \Gamma\left(\frac{p}{2}\right)}{2\pi^{\frac{p}{2}} \cdot r^p_{k(N)}} \cdot \frac{k(N)}{N+1} \qquad (7.8.5)$$

к истиной многомерной плотности $p(x_1, \ldots, x_p)$. Из (7.8.5) видно, что все вычислительные трудности оценивания многомерной плотности связаны только с нахождением радиуса $r_{k(N)} = |z - z_{(k(N))}|$ *.

Идея второго типа — поблочная обработка данных — пока (насколько известно автору) не приложена непосредственно к оцениванию распределений вероятностей. Однако экономия в вычислениях, получаемая таким способом, оценена Фьюстелом и Дэвиссоном [1]. Например, при модификации ранговых тестов с помощью поблочной обработки объем вычислений уменьшается с $O(N^2)$ до $O(mN)$, где m — объем подвыборки; при этом оказывается, что уже при небольших m (порядка 10—15) эффективность процедур весьма близка к наибольшей достижимой.

Третий подход к оцениванию многомерных распределений — использование итеративных процедур — освещен в литературе довольно подробно. Оказалось, что к настоящему времени наибольшая необходимость в непараметрических оценках многомерных распределений сложилась в той области приложений статистики, которая посвящена так называемым процедурам распознавания образов. Поэтому многие проблемы теории рекурсивных оценок распределений изложены в теории обучающихся или самообучающихся систем как частные, вспомогательные вопросы (см., например, Я. З. Цыпкин [1])**.

Структура итеративных оценок может быть описана следующим образом. Пусть, например, $f_k(z)$ — некоторая оценка функции $f(z)$, построенная на выборке объема k. Если при получении $(k+1)$-го выборочного значения z_{k+1} для нахождения $f_{k+1}(z)$ гребуется выполнение некоторой одинаковой

---

* Похожую идею, но с ориентацией на распознавание образов, развивают Честер [1], Себестьен и Эдл [1].

** Очевидно, итеративные оценки могут с успехом применяться и в одномерном случае; однако вычислительные достоинства таких оценок наиболее полно раскрываются при оценивании многомерных распределений.

для всех k совокупности операций L над $f_k(z)$ и z_{k+1}, то оценка называется итеративной:

$$f_{k+1}(z)=L[f_k(z),\ z_{k+1}]. \tag{7.8.6}$$

По-видимому, для большинства оценок распределений, рассмотренных в предыдущих параграфах, можно построить их итеративные аналоги. Например, итеративный вариант оценки (7.6.7) плотности разложением по весовым функциям выглядит как

$$p_N(z)=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{h_i^p}\cdot K\left(\frac{z-Z_i}{h_i}\right), \tag{7.8.7}$$

где z — p-мерный вектор; Z_i — p-мерная i-я выборочная точка; $\{h_i\}$ — последовательность убывающих положительных чисел, таких, что

$$\text{а) } 0\leqslant h_N\leqslant\ldots\leqslant h_1\leqslant 1;\quad \text{б) } \sum_{n=1}^{\infty}\frac{h_n}{n}<\infty;$$

$$\text{в) } \sum_{n=1}^{\infty}\frac{1}{n^2 h_n^m}<\infty; \tag{7.8.8}$$

а $K(z)$ — весовая функция p-мерного аргумента, обладающая свойствами плотности вероятностей, а также

$$\text{а) } K(z)\leqslant A<\infty \text{ для всех } z\in R^m;$$

$$\text{б) } \int \|z\|\,K(z)\,dz=B<\infty. \tag{7.8.9}$$

В этих предположениях можно доказать следующую теорему о сходимости оценки (7.8.7) к оцениваемой плотности (Вулвертон и Вагнер [1]):

Теорема 7. 8. 1. Если $p(z)$ удовлетворяет условиям Липшица, то с вероятностью единица

$$\lim_{N\to\infty}\int[p_N(z)-p(z)]^2\,dz=0.$$

Для другой оценки типа (7.7.2) или (7.7.11) построение рекуррентного аналога производится путем введения итеративных оценок коэффициентов разложения. Покажем, например, как строится оценка функции распределения $F(z)$ (см. Кашньяп и Блэйдон [1]).

Пусть функция $F(z)$ представима в виде ряда

$$F(z)=\sum_{k=1}^{\infty}\alpha_k^*\varphi_k(z)\doteq\sum_{k=1}^{m}\alpha_k^*\varphi_k(z)=\alpha^{*T}\varphi(z), \tag{7.8.10}$$

где α^* — m-компонентный вектор коэффициентов, $\{\varphi_k(z)\}$ — система функций, обладающих свойствами, необходимыми для

представимости $F(z)$ в виде (7.8.10) — в частности, матрица $E\{\varphi(z)\cdot\varphi^T(z)\}$ должна быть положительно определена.

Задача теперь сводится к нахождению итеративных оценок α коэффициентов α^*, которые бы минимизировали некоторый функционал, характеризующий точность получаемой оценки $\hat{F}_N(z)=\alpha^T\cdot\varphi(z)$. Пусть, например, требуется минимизировать среднеквадратичную ошибку.

$$I(\alpha)=E\{F(z)-\alpha^T\varphi(z)\}^2.$$

Значения $\alpha=\alpha^*$, минимизирующие $I(\alpha)$, равны

$$\alpha^*=\operatorname{Arg}\{\min_{\alpha} I(\alpha)\}=\frac{E\{\varphi(z)F(z)\}}{E\{\varphi(z)\varphi^T(z)\}}.$$

Определим итеративную оценку α для α^*, ограничившись первым приближением разложения α^* по приращениям величины $I(\alpha)$:

$$\alpha(i+1)=\alpha(i)-\rho\left.\frac{\partial I(\alpha)}{\partial\alpha}\right|_{\alpha=\alpha(i)}, \qquad (7.8.11)$$

где $\alpha(i)$ — оценка α^* на i-м шаге итерации. Так как $F(z)$ неизвестно, второй член тоже неизвестен. Предположим, что мы можем сконструировать некоторую функцию $y(\alpha, z)$ такую, что

$$E[y(\alpha, z)|\alpha]=\frac{\partial I(\alpha)}{\partial\alpha}.$$

Заменим этой функцией величину $\frac{\partial I(\alpha)}{\partial\alpha}$ в (7.8.11):

$$\alpha(i+1)=\alpha(i)-\rho(i)\cdot y(\alpha(i), z(i)). \qquad (7.8.12)$$

Требуя, чтобы (7.8.12) сходилось к α^* при $N\to\infty$, мы полагаем, что $\alpha(i)$ может быть произвольным; $z(i)=z_i$ — последнее выборочное значение; и $\rho(i)$ устремляется к нулю по мере увеличения N и сближения α и α^*. Как и ранее (см. § 7.7), непараметрическим фактом, обеспечивающим гребуемую сходимость оценки, служит сходимость эмпирической функции распределения (обозначим ее через $\hat{F}_N$, чтобы не путать с $\hat{F}_N$) к функции F; определим $y(\alpha, z)$ как

$$y(\alpha, z(i))=-\varphi(z(i))[\hat{F}_i-\alpha^T\varphi(z(i))]. \qquad (7.8.13)$$

Чтобы избежать необходимости запоминать все i выборочных значений $z_1, \ldots, z_i$, предлагается пользоваться «скользящей» оценкой $\hat{F}_i$, на n последних выборочных значениях:

$$\hat{F}_i=\frac{1}{n}\sum_{j=i-n}^{i-1} C(z-z_i),$$

дисперсия которой равна $\operatorname{Var}\hat{F}_i=F(x)[1-F(x)]/n\leqslant 1/4n$. Покажем, что $y(\alpha, z)$ обладает требуемым свойством:

$$E(y(\alpha, z)|\alpha)=-E\{\varphi(z(i)(\cdot\hat{F}_i\}+E\{\varphi(z)\varphi^T(z)\}\cdot\alpha=$$

$$=-E\{\varphi(z)\}\cdot F(z)+E\{\varphi(z)\varphi^T(z)\}\alpha=\frac{\partial I(\alpha)}{\partial\alpha}.$$

Итак, окончательно алгоритм (7.8.12) запишется в виде

$$\alpha(i+1)=\alpha(i)+\rho(i)\varphi(z(i))[\hat{F}_i(z)i))-\alpha^T\varphi(z(i))].$$

При $\rho(i)$, удовлетворяющих условиям

$$\lim_{i\to\infty}\rho(i)=0,\ \sum_{i=1}^{\infty}\rho(i)=\infty,\ \sum_{i=1}^{\infty}\rho^2(i)<\infty,$$

α сходится к α^* по вероятности и в среднеквадратичном смысле (Кашняп и Блэйдон [2]) *.

Подчеркнем недостатки только что рассмотренной оценки F_N функции F.

1. Даже при $N\to\infty$ она дает лишь приближенную оценку F из-за конечности m.

2. Максимально достижимая точность сильно зависит от удачного выбора системы функций для разложения неизвестной $F(z)$.

3. $F_N(z)$ может оказаться не обладающей всеми свойствами функции распределения: на краях области оценивания она может выйти за пределы [0, 1], а внутри области оценивания $F_N(z)$ может не быть везде неубывающей; хотя на практике это случается редко.

4. При большой размерности p объем вычислений остается большим. Остается надеяться, что могут быть предложены другие итеративные алгоритмы, обладающие более высокими качествами. Например, интересный подход к рассмотрению сходимости рекуррентных оценок типа (7.8.12) развивает Л. Дэвиссон [1]: он дает нижнюю оценку вероятности сходимости при невыполнении достаточных условий сходимости в среднеквадратичном и сходимости с вероятностью единица.

---

* Аналогичный алгоритм может быть построен для оценивания плотности; мы не будем его здесь воспроизводить, а интересующиеся могут найти его в упомянутой работе Кашньяпа и Блэйдона [1].

§ 7.9. КОНТУРНОЕ ОЦЕНИВАНИЕ ФУНКЦИИ РАСПРЕДЕЛЕНИЯ

Аналогично тому, как различаются точечные и интервальные оценки параметров, для функции распределения могут быть введены не только оценки, являющиеся одиночными линиями (в одномерном случае) или поверхностями (в многомерном), но и оценки, представляемые парой кривых или, соответственно, поверхностей, между которыми с заданной вероятностью заключена истинная функция распределения. В одномерном случае для этих двух типов оценок удобно ввести наименования соответственно «линейчатых» и «контурных» (за неимением лучших терминов эти же названия будем употреблять и в многомерном случае). Предыдущие параграфы были посвящены линейчатым оценкам; в данном параграфе обсудим некоторые вопросы контурного оценивания функций распределения.

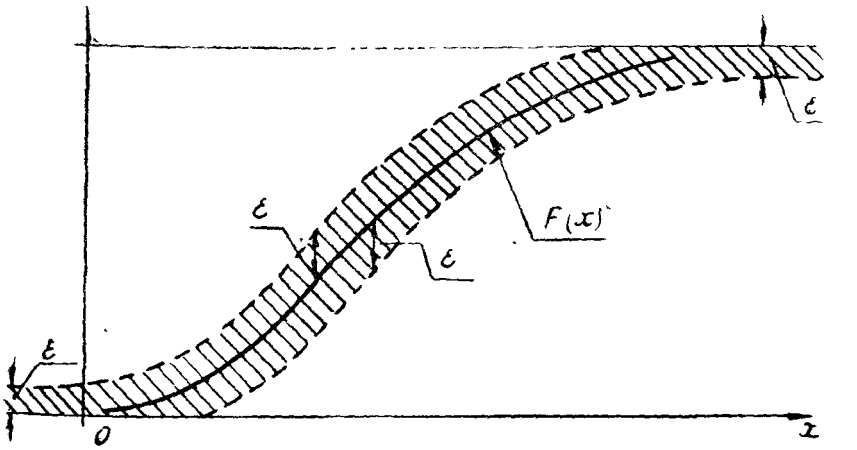

Рис. 7.9.1

Существуют два достаточно развитых подхода к получению контурных оценок для непрерывных $F(x)$. Первый восходит к идеям Колмогорова и Смирнова и состоит в том, чтобы указать контуры кривых, отстоящих на заданном расстоянии ε от $F(x)$ (см. рис. 7.9.1), задав тем самым полосу (заштрихованную на рисунке), в которой находится $F(x)$. Если $F(x)$ задано, то можно показать, что вероятность $P_N(\varepsilon)$ того, что модуль разности между $F(x)$ и эмпирической функцией распределения $F_N(x)$ (см. § 7.2) превысит ε при истинном распределении $F(x)$, не зависит от $F(x)$. Асимптотическое распределение $P_N(\varepsilon)$ было получено и табулировано Смирновым [1, 2], явные выражения для $P_N(\varepsilon)$ (для односторон-

него контура) получены Бирнбаумом и Тинги [1]. Особенность этого подхода состоит в необходимости задавать до опыта «опорное», гипотетическое распределение $F(x)$. Тем самым метод фактически сводится не столько к оцениванию неизвестного распределения $F(x)$, сколько к проверке гипотезы о том, лежит ли неизвестное распределение в заданной указанным выше способом полосе. Эта задача относится к классу задач проверки согласия, которые мы рассмотрим отдельно в главе 9.

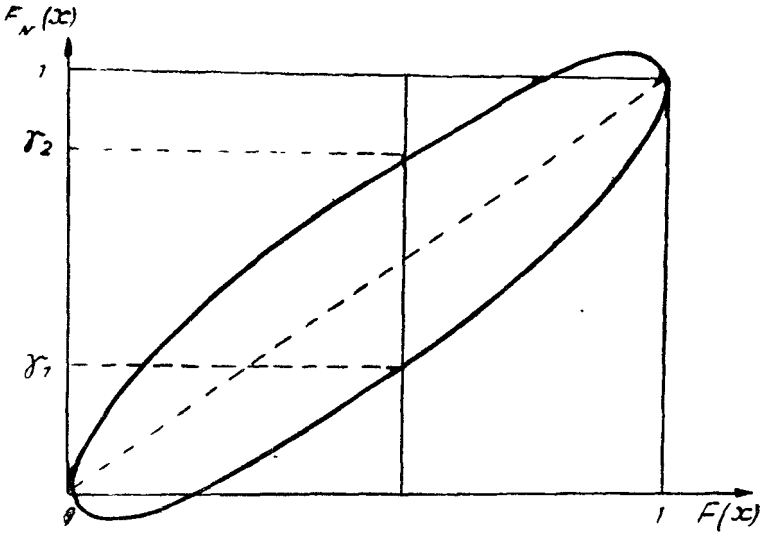

Рис. 7.9.2

Второй подход к контурному оцениванию функции распределения базируется на том, что эмпирическая функция распределения есть оценка вероятности относительной частотой; так как последняя асимптотически нормальна со средним $F(x)$ и дисперсией $\dfrac{F(x)[1-F(x)]}{N}$, где $F(x)$ — искомая вероятность, то для каждого x можно найти доверительный интервал и тем самым — контурную оценку для $F(x)$. Считая, что распределение в точности нормально, и задав границы доверительного интервала γ_1 и γ_2 соотношением $F(x) \pm \pm\lambda\sqrt{F(x)[1-F(x)]/N}$, получим, что $F(x)$ лежит в пределах

$$\frac{N}{N+\lambda^2}\left[F_N(x)+\frac{\lambda^2}{2N} \pm \sqrt{\frac{F_N(x)[1-F_N(x)]}{N}+\frac{\lambda^2}{4N^2}}\right]$$

(см. рис. 7.9.2, Г. Крамер [1]). Как видим, ширина довери-

тельного интервала убывает по мере удаления к хвостам распределения. Это приводит к мысли задавать границы для контурной оценки распределения $F(x)$ не через эллипс, соответствующий рисунку 7.9.2, а более простыми кривыми или ломаными, отобразив лишь указанную общую тенденцию. Вопросы вычисления доверительных вероятностей для областей, получаемых таким образом, рассматривали Андерсон и Дарлинг [1], С. Малмквист [1] и др.

§ 7.10. НЕСМЕЩЕННЫЕ И СОСТОЯТЕЛЬНЫЕ ОЦЕНКИ РАСПРЕДЕЛЕНИЙ ВЫСШИХ ПОРЯДКОВ ПО ОДНОМЕРНОЙ ВЫБОРКЕ

В ряде случаев возникает необходимость построения оценки функции распределения более высокого порядка, $F(x^{(1)}, \ldots, x^{(k)})$. Такая задача имеет место как для случая зависимых выборочных значений, так и для независимых $x_1, \ldots, x_N$. Казалось бы, в последнем случае можно воспользоваться произведением одномерных эмпирических функций распределения. Однако легко показать, что такая оценка не будет являться несмещенной, что часто нежелательно.

Покажем, как по одномерной выборке $x_1, \ldots, x_N$ можно построить несмещенную и состоятельную оценку распределения $F(x^{(1)}, \ldots, x^{(k)})$ (Симахин, Тарасенко, Шуленин [1]).

Теорема 7.10.1. Пусть $x_1, \ldots, x_N$ — выборка независимых и одинаково распределенных случайных переменных. Определим

$$F_N(x^{(1)}, \ldots, x^{(k)}) = \frac{1}{\binom{N}{k} k!} \sum_{M} \prod_{j=1}^{k} C(x^{(j)} - x_{i_j}), \qquad (7.10.1)$$

где $C(y) = \{1$ при $y \geqslant 0$; 0 при $y < 0\}$, и M — совокупность индексов $\{i_j\}$ такая, что ни один из них не совпадает с другими. Тогда F_N является несмещенной и состоятельной оценкой $F = \prod_j F(x^{(j)})$.

Доказательство. Несмещенность F_N очевидна. Для вычисления дисперсии $D[F_N]$ необходимо сначала найти

$$F_N^2 = \left[\binom{N}{k} k!\right]^{-2} \left[\sum_{M} \prod_{j} C(x^{(j)} - x_{i_j})\right]^2.$$

При возведении в квадрат суммы и последующем усреднении получатся члены, пропорциональные следующим выражениям: $F^2 = F_{(1)}^2 \ldots F_{(k)}^2$; $F_{(1)} F_{(2)}^2 \ldots F_{(k-1)}^2$; $F_{(1)} F_{(2)} F_{(3)}^2 \ldots F_{(k-2)}^2$; $\ldots$; $F_{(1)} \ldots F_{(k)}$

где $F_{(j)}=F(x^{(j)})$. Так как $DF_N=EF_N^2-E^2F_N=EF_N^2-F^2$, то нас прежде всего интересуют члены, пропорциональные F^2. Эти члены входят в сумму вида $\sum\sum\prod_{M\,I\,J} C_{i_j}^{(j)}\cdot C_{i_e}^{(j)}$, в которой ни один из индексов в множествах M и L не совпадает с другими. Таких членов будет $\binom{N}{2k}\cdot(2k)!$ Поэтому после вычитания F^2 останется член, равный $\left[\binom{N}{2k}\cdot(2k)!-1\right]F^2$. Число всех остальных членов (поскольку всех членов $\left[\binom{N}{k}k!\right]^2$) равно $\left[\binom{N}{k}k!\right]^2-\left[\binom{N}{2k}(2k)!-1\right]$. Так как $F_{(j)}\leqslant 1$, то $DF_N\leqslant$ $\leqslant\left[\binom{N}{k}k!\right]^{-2}\times\left\{\left[\binom{N}{k}k!\right]^2-\left[\binom{N}{2k}(2k)!\right]+1\right\}$, что дает $DF_N\sim$ $\sim 0(N^{-1})$, откуда и следует утверждение теоремы.

При наличии зависимости между выборочными значениями соответствующую теорему удалось доказать лишь для случая m-зависимого процесса. Напомним (Фрейзер [1]), что m-зависимым процессом называется процесс с дискретным временем, значения x_l и x_l которого для всех i и l, таких, что $|i-l|>m$, статистически независимы. Иначе говоря, если изъять из выборки m примыкающих друг к другу значений, то каждое из значений справа от изъятых будет независимым от любого из значений слева. В силу зависимости между выборочными значениями мы лишены возможности осуществлять произвольные перестановки, как это имело место в случае независимой выборки.

Теорема 7.10.2. Пусть x_1, ..., x_N — зависимая выборка из реализации стационарного эргодического m-зависимого процесса. Определим

$$F_N^*(x^{(t_1)},\ldots,x^{(t_k)})=\frac{1}{N-t_k}\sum_{i=1}^{N-t_k}\prod_{t_j=1}^{k}C(x^{(t_j)}-x_{i+t_j}),\quad(7.10.2)$$

где $t_1\leqslant\ldots\leqslant t_k\leqslant m$ — целые числа, $C(y)$ определена выше. Тогда F_N^* является несмещенной и состоятельной оценкой

$$F=F(x^{(t_1)},\ldots,x^{(t_k)}).$$

Доказательство. Несмещенность доказывается непосредственным усреднением (7.10.2). Для доказательства состоятельности вычислим дисперсию F_N^*. Учитывая, что $C^2(y)=$ $=C(y)$, имеем:

$$EF_N^{*2}=\frac{1}{(N-t_k)^2}E\left[\sum_{i=1}^{N-t_k}\prod_{t_j=1}^{k}C(x^{(t_j)}-x_{i+t_j})\right]^2=$$

$$= \frac{1}{(N-t_k)^2}\left\{\sum_{i=1}^{N-t_k} E \prod_{t_j=1}^{k} C(x^{(t_j)}-x_{i+t_j}) + \right.$$

$$\left. + \sum_{i \neq l} E \prod_{t_j=1}^{k} C(x^{(t_j)}-x_{i+t_j})\, C(x^{(t_j)}-x_{l+t_j}) \right\}. \qquad (7.10.3)$$

Первая сумма в (7.10.3) равна $(N-t_k)F$. Поскольку выборка взята из m-зависимого процесса, то члены второй суммы, для которых $|i-l|>m$, дадут при усреднении F^2. Найдем сначала число n зависимых членов во второй сумме. В интервале длиной m, начинающемся с x_1, содержится $\binom{m}{2}$ зависимых членов. При сдвиге на один шаг появится $(m-1)$ новых зависимых пар; таких шагов можно сделать $(N-m+1)$. Следовательно, $n=\binom{m}{2}+(N-m+1)(m-1)$. Так как общее число R членов в сумме равно $(N-t_k)^2-(N-t_k)$, то число членов, пропорциональных F^2, равно $R-n$. Отсюда, дисперсия F_N^* равна

$$DF_N^* \leqslant \frac{1}{N-t_k}F - \frac{1}{(N-t_k)^2}\left[(N-t_k)+\binom{m}{2}+(N-m+1)(m-1)\right]F^2 +$$

$$+ \frac{1}{(N-t_k)^2}\left[\binom{m}{2}+(N-m+1)(m-1)\right]F_0 \leqslant \frac{F_0}{N-t_k}, \qquad (7.10.4)$$

где

$$F_0 = \max\left\{\sup_x [F(x^{(1)},\ldots, x^{(t_k-1)})\cdot F_1;\ F(x^{(1)},\ldots, x^{(t_k-2)})\cdot F_1^2;\ldots\right.$$

$$\left.\ldots;\ F(x^{(1)}, x^{(2)})\cdot F_1^{t_k-1}]\right\}.$$

Очевидно, что (7.10.4) стремится к нулю при $N\to\infty$, что и завершает доказательство теоремы.

Возможно обобщение теоремы 7.10.2 на случай процессов, подчиняющихся условию сильного перемешивания, что требует более тонких методов (Тарасенко [8]).

ГЛАВА VIII

НЕПАРАМЕТРИЧЕСКОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ

§ 8.1. О НЕПАРАМЕТРИЧЕСКОМ ПОДХОДЕ К ЗАДАЧЕ СТАТИСТИЧЕСКОГО ОЦЕНИВАНИЯ

Задача оценивания параметра возникает в статистике в связи с изучением степени влияния некоторого фактора на статистические свойства измеряемой случайной величины.

Экспериментатор, знающий природу изучаемого фактора (или выдвигающий предположение о его природе), должен указать, какой параметр распределения будет изменяться под влиянием данного фактора. Например, физик утверждает, что появление полезного сигнала в шуме приводит к увеличению средней мощности принимаемого сигнала; инженер ожидает, что предложенное им усовершенствование технологии обеспечит уменьшение допусков изготовляемой детали; тренер считает, что его методика тренировок позволит повысить средние результаты атлетов; агрохимик надеется, что синтезированное им удобрение увеличит урожайность данной культуры, и т. д. и т. п. После того, как «подозреваемый» параметр указан, начинается работа статистика, задачей которого является построение процедуры (по возможности — наилучшей) для оценки этого параметра. Важно подчеркнуть, что решение задачи будет более успешным, если экспериментатор укажет не только единственный (хотя, возможно, и наиболее существенный) параметр, не только еще один-два параметра, которые тоже могут изменяться (пусть даже гораздо слабее, чем основной) под влиянием тех же факторов, но, если это возможно, укажет и параметры, которые определенно не изменятся. Всякие сведения этого сорта могут оказаться полезными для повышения качества получаемых оценок.

Влияние любого фактора может быть обнаружено и оценено статистическими методами лишь в том случае, если оно приводит к какому-нибудь изменению распределения наблюдаемой случайной величины. Параметры, рассматриваемые статистикой, являются количественными характеристиками тех или иных изменений распределения и, следовательно, выражаются некоторым известным образом через эти распределения. Иначе говоря, *параметры, подлежащие статистическому оцениванию, всегда являются известными функционалами от распределения выборочных значений* (или известными функциями таких функционалов). Разница между параметрическими и непараметрическими методами оценивания параметров* та, что при построении параметрических оценок используется знание типа функции распределения, а в непараметрическом случае мы лишены такой возможности.

Казалось бы, поскольку *любые* параметры распределения могут рассматриваться как функционалы этих распределений, то разница между параметрическим и непараметрическим подходами к оцениванию лежит в чисто техниче-

* Неуклюжесть и семантическая противоречивость этого выражения лишний раз иллюстрируют недостатки термина «непараметрический», обсужденные в главе 1.

ской, вычислительной области. Такое неверное понимание может вызвать недоразумения и даже заблуждения. Например, довольно распространено оценивание параметров по методу моментов. Можно было бы подумать, что поскольку в процессе оценивания используются только моменты, данный метод обладает непараметричностью. Однако конкретная связь между оцениваемым параметром и моментами однозначно определяется видом распределения. Поэтому если истинное распределение хотя бы «незначительно» отличается от принятого в модели, то оценка, получаемая по методу моментов, может содержать дополнительную неконтролируемую погрешность Метод моментов — один из параметрических методов, для которых расхождение между истинным и предполагаемым распределениями теоретически недопустимо.

Оценки параметров будем называть непараметрическими, если они получены без конкретизации вида распределения. Акцент, таким образом, делается на методах построения оценок; вопрос о наличии или отсутствии зависимости качества таких оценок от вида распределения должен рассматриваться каждый раз отдельно. Интересно, что непараметрическое оценивание дает оценки, которые по качеству часто не уступают параметрическим оценкам. Однако гораздо более важным является тот факт, что непараметрические оценки превосходят по качеству параметрические в том случае, когда истинное распределение $G(x)$ отличается от положенного в основу параметрической оценки. Можно поэтому высказать рекомендацию: если вы не очень уверены в том, что измеряемая величина действительно подчиняется известному распределению, то лучше пользоваться непараметрическими оценками.

Можно усмотреть два разных способа задания параметров, характеризуемых функционалами распределений: явное и неявное. В первом случае параметр θ задается в виде

$$\theta_0 = J[F(x)], \tag{8.1.1}$$

где J — известный функционал, $F(x)$ — неизвестная функция распределения. Ко второму способу прибегают, когда по каким-то причинам θ нельзя определить в виде (8.1.1), но из самой природы задачи следует, что оцениваемый параметр связан с некоторыми особенностями известного функционала J Например, иногда удается определить функционал J, о связи которого с θ известно лишь то, что он (для любых распределений F из некоторого класса Ω) достигает экстремума при истинном значении параметра $\theta = \theta_0$:

$$\theta_0 = \{\theta : \operatorname{extr} J[F(x); \theta]; F \in \Omega\}. \tag{8.1.2}$$

Некоторые особенности возникают при определении функционалов на некотором множестве классов распределений $F_i \in \Omega_i$, $i=1, 2, \ldots, k$, что соответствует многовыборочным задачам оценивания.

В данной главе мы приведем ряд результатов по непараметрическому оцениванию параметров. Пока еще трудно говорить о какой-либо единой теории непараметрического оценивания; скорее, речь может идти о совокупности методов построения непараметрических оценок. Если же попытаться классифицировать конкретные объекты непараметрического оценивания, то похоже, что все параметры можно разделить на четыре типа:

а) локальные характеристики функции распределения (квантили, экстремумы плотности);

б) линейные функционалы (т. е. математические ожидания известных функций случайной переменной) и их степени;

в) нелинейные функционалы распределений;

г) параметры, известным образом входящие в неизвестное распределение.

Случаи б) и в) соответствуют различным конкретизациям функционала (8.1.1); случай г) связан с (8.1.2). Для оценивания каждого из четырех типов параметров предложено несколько разных методов; эти методы будут изложены ниже.

§ 8.2. ТОЧЕЧНОЕ И ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ КВАНТИЛЕЙ

Квантиль уровня α есть значение x, удовлетворяющее уравнению $F(x)=\alpha$, или $F^{-1}(\alpha)=x$. По положению квантилей того или иного уровня можно делать различные суждения о характере случайной величины.

Благодаря непараметрическому факту сходимости порядковой статистики x_R к квантилю уровня $R/(N+1)$ (см. § 3.5) порядковые статистики являются естественно напрашивающимися точечными оценками соответствующих квантилей:

$$\hat{x}_\alpha = F^{-1}\left(\alpha = \frac{R}{N+1}\right) = x_{(R)}. \qquad (8.2.1)$$

Эти удобные и широко применяемые оценки обладают, однако, рядом недостатков. Во-первых, при этом можно оценивать лишь квантили с определенными уровнями, кратными $1/(N+1)$. Во-вторых, точность этих оценок характеризуется зависимостью их стандартного отклонения от объема

выборки порядка $O(N^{-\frac{1}{2}})$; желательно было бы уменьшить и вклад смещения. В-третьих, по мере приближения к краям упорядоченной статистики точность этих оценок убывает, так что квантили малых и больших уровней лучше не оценивать данным способом.

Поэтому значительный интерес представляет исследование Дж. Уолша [2] оценок квантилей *произвольного* уровня в виде линейной комбинации заданного числа m порядковых статистик:

$$\hat{x}_{\alpha m}=\sum_{i=1}^{m} a_i x_{(R_i)}. \qquad (8.2.2)$$

Из этой работы Уолша следует, что при соответствующем выборе коэффициентов $\{a_i\}$ и номеров $\{R_i\}$ порядковых статистик, можно добиться, чтобы

$$E\hat{x}_{\alpha m}=x_\alpha+0(N^{-\frac{m}{2}}), \qquad (8.2.3)$$

если $F(x)$ имеет достаточное число производных в точке x_α, что не является на практике слишком жестким ограничением.

Если не вдаваться в подробности, то схематично метод Уолша, состоящий из последовательности определенных операций, можно проиллюстрировать следующим образом. Сначала задаются числа m, N и α. Метод предназначен для получения оценок, удовлетворяющих (8.2.3); и не любые комбинации $\{a_i\}$ и $\{R_i\}$ могут удовлетворить этому соотношению. Например, если $m=3$, $N=168$, $\alpha=0{,}28$, то кроме $x_{(47)}$, ближайшего к $x_{0,28}$ выборочного квантиля, необходимо взять еще две порядковых статистики, номера которых отличаются от 47 на ближайшее целое к величине k,

$$K\geqslant\sqrt{N+1}\,[10\prod_{i>j-1}^{m}(i-j)]^{-\frac{1}{m}},$$

которая в нашем примере равна 4,8. Таким образом, искомая оценка будет иметь вид:

$$x_{0,28}=a_1 x_{(42)}+a_2 x_{(47)}+a_3 x_{(52)}.$$

Решая специальную систему уравнений для величин $\{a_i\}$, Уолш окончательно получает:

$$x_{0,28}=-0{,}7113\cdot x_{(42)}+2{,}3587\cdot x_{(47)}-0{,}6474\cdot x_{(52)}.$$

Однако ограничения этого метода допускают не всякие комбинации чисел α, N и m. Например, при $N=80$ и $\alpha=0{,}91$ значения m от 2 до 7 не дают оценок, удовлетворяющих (8.2.3).

Теория точечных оценок квантилей еще не приобрела завершенного вида. Некоторые ее аспекты освещены Лойнсом [1], но многие вопросы остаются открытыми. Например, представляло бы интерес исследование поведения оценок вида (8.2.2) без требования (8.2.3); сравнение оценок квантилей по различным оценкам функции распределения и т. д.

Гораздо более подробно изучены свойства интервальных оценок квантилей. (См., например, С. Уилкс [1]). Возьмем две произвольных порядковых статистики $x_{(k_1)}$ и $x_{(k_1+k_2)}$ и определим вероятность того, что квантиль x_α уровня α находится между ними: $Pr\{x_{(k_1)} \leqslant x_\alpha \leqslant x_{(k_1+k_2)}\}$. В случае монотонности функции распределения эта вероятность равна:

$$Pr\{F(x_{(k_1)}) \leqslant \alpha \leqslant F(x_{(k_1+k_2)})\},$$

где $F(x)$ — функция распределения, из которого взята выборка. Но величины $u=F(x_{(k_1)})$ и $v=F(x_{(k_1+k_2)})$ являются соответствующими порядковыми статистиками выборки из равномерного распределения, и их совместное распределение известно:

$$f(u, v)=\frac{N!}{(k_1-1)!\,(k_2-1)!\,(N-k_1-k_2)!}\,u^{k_1-1}(v-u)^{k_2-1}\times \\ \times(1-v)^{N-k_1-k_2}. \qquad (8.2.4)$$

Отсюда, искомая вероятность выразится величиной

$$Pr\{x_{(k_1)} \leqslant x_\alpha \leqslant x_{(k_1+k_2)}\}=\int_\alpha^1\int_0^\alpha f(u, v)\,du\,dv= \\ =I_\alpha(k_1, N-k_1+1)-I_\alpha(k_1+k_2, N-k_1-k_2+1), \qquad (8.2.5)$$

где $I_\alpha(n_1, n_2)$ — неполная бета-функция,

$$I_\alpha(n_1, n_2)=\frac{\Gamma(n_1+n_2)}{\Gamma(n_1)\,\Gamma(n_2)}\int_0^\alpha x^{n_1-1}(1-x)^{n_2-1}\,dx.$$

Отметим теперь несколько особенностей интервальной оценки $[x_{(k_1)}, x_{(k_1+k_2)}]$ квантиля x_α.

Во-первых, найденная вероятность (8.2.5) того, что квантиль x_α окажется в данном интервале, не зависит от $F(x)$, поэтому этот интервал является доверительным интервалом для x_α при доверительной вероятности, равной правой части (8.2.5). Во-вторых, поскольку k_1 и k_2 выбраны произвольно, по выборке объема N можно построить $N(N-1)/2$ различных интервальных оценок для данного квантиля x_α. Поэтому возникает задача выбора «наилучшего» из этих интервалов. Естественно, что из числа интервалов, обладающих доверительной вероятностью, не меньшей заданной, обычно выби-

рается тот, для которого длина интервала минимальна, либо тот, для которого минимальна величина k_2.

Следует отметить работы Вайса [1], а также Блюма и Розенблатта [1], в которых рассматриваются вопросы нахождения доверительных интервалов для квантилей в тех случаях, когда заранее задаются требуемая длина интервала и доверительная вероятность.

§ 8.3. ОЦЕНИВАНИЕ ЛИНЕЙНЫХ ФУНКЦИОНАЛОВ U-СТАТИСТИКАМИ

В тех случаях, когда интересующий нас параметр θ является математическим ожиданием известной функции $\varphi(x)$ случайной величины X, распределение $G(x)$ которой неизвестно, для оценивания θ часто можно воспользоваться результатами теории так называемых U-статистик, краткому изложению которой и посвящен данный параграф*. Речь пойдет о том, каким образом оценить величину $\theta = E\varphi(x) = \int \varphi(x)\,dG(x)$, если φ задана, а G неизвестно, и какими свойствами будет обладать такая оценка.

Пусть на пространстве возможных значений случайной величины X определен класс Ω возможных распределений $\{G\}$. Будем предполагать далее, что выборка $x_1, x_2, \ldots, x_N$ состоит из независимых и одинаково распределенных значений (не обязательно одномерных), так что на пространстве X^N выборок определен класс распределений $\{\prod_{i=1}^{N} G(x_i)\}$, $G(x_i) \in \Omega$.

Вещественный параметр θ называется допускающим несмещенную оценку, или регулярным, если существует такая статистика $f(x_1, x_2, \ldots, x_N)$**, что

$$E[f(x_1, \ldots, x_N)] = \int_{X^N} f(x_1, \ldots, x_N) \prod_{i=1}^{N} dG(x_i) = \theta \qquad (8.3.1)$$

для всех $G \in \Omega$. Наименьший объем выборки, для которого параметр θ имеет несмещенную оценку, называется степенью m регулярного параметра; m является минимальным значением N, для которого выполняется равенство (8.3.1). Любую несмещенную оценку параметра θ, основанную на выборке объема m, будем называть ядром. Легко показать, что всегда существует симметричное ядро. Дейст-

* U-статистики были введены в 1948 г. Хёдингом [2]; детальное изложение теории U-статистик имеется в книге Фрейзера [1].

** Вопрос о том, каким образом связаны функции φ и f, будет обсужден позднее.

вительно, если $f(x_1, x_2, ..., x_m)$ несимметричная функция, то всегда можно построить симметричную функцию

$$f_S(x_1,..., x_m)=\frac{1}{m!}\sum_{P} f(x_{i_1}, x_{i_2},..., x_{i_m}), \qquad (8.3.2)$$

где суммирование проводится по всем перестановкам $(i_1, i_2, ..., i_m)$ чисел $(1, 2, ..., m)$. Так как все $f(x_{i_1}, x_{i_2},..., x_{i_m})$ — ядра, то f_S тоже является ядром.

Интересно отметить некоторые свойства регулярных параметров. Если θ_1 и θ_2 — регулярны и имеют соответственно степени m_1 и m_2, то $\theta_1+\theta_2$ и $\theta_1\cdot\theta_2$ тоже регулярны, и их степени меньше или равны соответственно $m=\max(m_1, m_2)$ и m_1+m_2. Действительно, если f_i является ядром для θ_i, то

$$\int_{X^m}[f_1(x_1,..., x_{m_1})+f_2(x_1,..., x_{m_2})]\prod_{i=1}^{m} dG(x_i)=\theta_1+\theta_2; \qquad (8.3.3)$$

$$\int_{X^{m_1+m_2}} f_1(x_1,..., x_{m_1}) f_2(x_{m_1+1},...$$

$$..., x_{m_1+m_2})\prod_{i=1}^{m_1+m_2} dG(x_i)=\theta_1\theta_2. \qquad (8.3.4)$$

Очевидно, справедливо и более общее утверждение, что любой полином от регулярных параметров является регулярным параметром. Может возникнуть вопрос, чем определяется степень m регулярного параметра. Если оцениваемый параметр выражается в виде полинома от математического ожидания известной функции, то m и есть как раз степень этого полинома. Случаи более сложной зависимости между нужным нам параметром и некоторым математическим ожиданием требуют особого рассмотрения.

Перейдем теперь к вопросу о построении оценки параметра по выборке объема $N>m$. Для любого ядра $f(x_1, x_2, ..., x_m)$ определим U-статистику выборки объема соотношением:

$$U(x_1,..., x_N)=\frac{1}{\binom{N}{m}}\sum_{C} f_S(x_{i_1},..., x_{i_m}), \qquad (8.3.5)$$

где суммирование ведется по всем $\binom{N}{m}$ сочетаниям m чисел $(i_1, ..., i_m)$ из $(1, 2, ..., N)$, а f_S — симметризованное ядро (8.3.2), соответствующее заданному ядру f. Ясно, что U-статистику можно также записать в форме

$$U(x_1,\ldots,x_N)=\frac{(N-m)!}{N!}\sum_P f(x_{i_1},\ldots,x_{i_m}), \qquad (8.3.6)$$

где суммирование ведется по всем перестановкам всех комбинаций из m чисел, которые можно выбрать из чисел $(1, 2, \ldots, N)$. Из (8.3.6) видно, что U-статистика является просто симметризованной по всей выборке формой ядра f.

Перечислим теперь в виде теорем* важнейшие свойства U-статистик.

Теорема 8.3.1. U-статистика является несмещенной оценкой параметра θ.

Доказательство легко следует из (8.3.5), (8.3.2) и (8.3.1).

Теорема 8.3.2. Если $f(x_1, \ldots, x_N)$ является несмещенной оценкой для θ и имеет дисперсию σ_f^2, то дисперсия σ_U^2 соответствующей U-статистики меньше или равна σ_f^2, причем равенство имеет место лишь в том случае, когда $f=U$ почти везде по мере $\prod_\iota dG$**.

Теорема 8.3.3. Если класс распределений $\{\prod_\iota dG\}$ обладает свойством полноты***, то U-статистика, соответствующая заданному регулярному параметру, существенно единственна и имеет наименьшую дисперсию среди несмещенных оценок этого параметра.

Теорема 8.3.4. Если $f(x_1, \ldots, x_N)$ — несмещенная оценка параметра θ, имеющая конечный второй момент, и $U_N(x_1, \ldots, x_N)$ — соответствующая ей U-статистика, то величина $\sqrt{N}(U_N-\theta)$ распределена асимптотически нормально с нулевым средним и конечной дисперсией.

Доказательство базируется на аддитивной структуре U-статистики и центральной предельной теореме.

Теорема 8.3.5. U-статистика, построенная на несмещенной оценке f параметра θ, имеющей конечный второй момент, является состоятельной оценкой θ.

Доказательство. Поскольку

$$Pr\{\theta-\varepsilon\leqslant U_N\leqslant\theta+\varepsilon\}=Pr\{-\sqrt{N}\varepsilon\leqslant\sqrt{N}(U_N-\theta)\leqslant\sqrt{N}\varepsilon\}, \qquad (8.3.7)$$

и $\sqrt{N}(U_N-\theta)$ имеет предельное нормальное распределение с конечной дисперсией (теорема 8.3.4), то при $N\to\infty$ и

* Некоторые из этих теорем приводятся без доказательств, так как в противном случае пришлось бы резко увеличить объем параграфа: опущенные доказательства (которые можно найти у Фрейзера в [1]) основываются на ряде результатов, не содержащихся в данной книге.

** Аналогичное утверждение справедливо и для средних рисков при любых строго выпуклых функциях потерь.

*** Требование полноты сводится, по существу, к тому, чтобы не существовало несмещенной оценки $h(x_1, \ldots, x_N)$ нуля кроме тривиальной (т. е. равной нулю почти везде).

$\sqrt{N}\varepsilon\to\infty$ правая часть (8.3.7) стремится к единице, т е. U-статистика сходится к θ по вероятности.

Перечисленные выше свойства U-статистик свидетельствуют об их высоких качествах, благодаря которым они играют существенную роль в непараметрическом оценивании.

Перейдем теперь к вопросу о том, как строить ядра U-статистик по заданной функции $\varphi(x)$. Продемонстрируем эту процедуру на ряде примеров с тем, чтобы общий рецепт сформулировать впоследствии.

Первые несколько примеров относятся к выборкам N независимых измерений случайной величины X, непрерывно распределенной на числовой оси.

Пример 1. Первый момент μ_1 является регулярным параметром, поскольку он может быть представлен в виде:

$$\theta=\mu_1=EX=\int xg(x)\,dx.$$

Функция $\varphi(x)$ равна x. Минимальный объем выборки, дающий несмещенную оценку θ, равен 1, поскольку $EX_1=\mu_1$. Следовательно, μ_1 имеет степень 1 и $f(x_1)=x_1$. Соответствующая U-статистика равна $\frac{1}{N}\sum x_i$; т. е. выборочному среднему.

Пример 2. Квадрат первого момента $\theta=\mu_1^2$ является степенью регулярного параметра, следовательно, он регулярен. Так как $\mu_1^2=[EX]^2=EX_1X_2$, следовательно, степень параметра θ равна 2 и $f(x_1, x_2)=x_1x_2$. Соответствующая U-статистика равна $[N(N-1)]^{-1}\sum\limits_{i\neq j}x_ix_j$.

Пример 3. Дисперсия σ^2 может быть записана в виде полинома регулярных параметров, следовательно, она тоже регулярна:

$$\theta=\sigma^2=E[(X-\mu_1)^2]=EX^2-[EX]^2.$$

В силу предыдущих двух примеров степень σ^2 равна 2 и $f(x_1, x_2)=x_1^2-x_1\cdot x_2$. Симметризация ядра дает: $f_S(x_1, x_2)=\frac{1}{2}(x_1^2-x_1x_2+x_2^2-x_2x_1)=(x_1-x_2)^2/2$. Соответствующая U-статистика равна

$$U(x_1,\ldots, x_N)=\frac{2}{N(N-1)}\sum_{i<j}\frac{(x_i-x_j)^2}{2},$$

что после перегруппировки и приведения подобных приводит к выражению для выборочной дисперсии:

$$\frac{2}{N(N-1)}\sum_{i<j}\frac{(x_i-x_j)^2}{2}=\frac{1}{N(N-1)}\{(N-1)\sum x_i^2-2\sum x_ix_j\}=$$

$$=\frac{1}{N}\sum x_i^2-\frac{1}{N(N-1)}[(\sum x_i)^2-\sum x_i^2]=\frac{1}{N-1}\sum x_i^2-$$

$$-\frac{1}{N(N-1)}\left[\sum x_i^2-\frac{1}{N}(\sum x_i)^2\right]=\frac{1}{N-1}(\sum x_i^2-N\mu_1^2).$$

По аналогии с примерами 1—3 строятся U-статистики для начальных и центральных моментов высших порядков и для кумулянтов распределения (конечно, порядок оцениваемых моментов не может быть выше объема выборки).

Пример 4. Вероятность события $a\leqslant x\leqslant b$ может быть выражена в виде:

$$\theta=P_G(a, b)=E[\pi_{ab}(x)]=\int_{-\infty}^{\infty}\pi_{ab}(x)g(x)\,dx,$$

где

$$\pi_{ab}(x)=\begin{cases}1, & X\in[a, b],\\ 0, & X\in[a, b].\end{cases}$$

Параметр θ является регулярным со степенью 1, и $f(x_1)=\pi_{ab}(x_1)$. Соответствующая U-статистика равна

$$U=\frac{1}{N}\sum_i\pi_{ab}(x_i),$$

т. е. является относительной частотой, и обладает наименьшей дисперсией среди несмещенных оценок вероятности.

Заметим, что результаты примеров 1—4 остаются справедливыми и для дискретных распределений.

Приведем теперь примеры построения U-статистик для оценок параметров в двумерном случае. Пусть имеется независимая двумерная выборка (x_1, y_1), (x_2, y_2), ..., (x_N, y_N). Нахождение первичных ядер производится аналогично тому, как это делается в одномерном случае; однако симметризация ядер и построение U-статистик существенно определяются тем, являются ли переменные X и Y независимыми внутри каждой пары (x_i, y_i). В случае независимости допустима (и необходима) симметризация по каждой переменной в отдельности, а при наличии зависимости симметризация не должна разрывать связанных пар.

Пример 5. Пусть оценивается параметр $\theta=E(X\cdot Y)$. Очевидно, несмещенной оценкой θ с минимальным объемом выборки будет $f(x_1, y_1)=x_1y_1$. Дальнейшая процедура построения U-статистики будет различной для зависимых и независимых X и Y.

а. Случай независимых X и Y. Симметризуя $f(x_1, y_1)$ сначала по x, имеем:

$$f_{S_x}=\left(\frac{1}{N}\sum x_i\right)\cdot y_1.$$

Производя далее симметризацию по y, окончательно получаем:

$$U_H=\left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)\left(\frac{1}{N}\sum_{i=1}^{N}y_i\right).$$

б. Случай зависимых X и Y. Теперь $f(x_1, y_1)$ следует симметризовать, не разрывая статистически связанных переменных:

$$U_3=\frac{1}{N}\sum_{i=1}^{N}x_i y_i.$$

Пример 6. Пусть по выборке $(x_1, y_1), \ldots, (x_N, y_N)$ оценивается дисперсия суммы X и Y, $\theta=\sigma^2(X+Y)$. Ядро минимального объема 2 равно

$$f[(x_1, y_1), (x_2, y_2)]=\frac{1}{2}[(x_1+y_1)-(x_2+y_2)]^2.$$

а. X и Y независимы. Аналогично примеру 3 получаем:

$$U_H=\frac{1}{N-1}\sum_{i=1}^{N}(x_i^2-\mu_x)^2+\frac{1}{N-1}\sum_{j=1}^{N}(y_j^2-\mu_y)^2.$$

б. X и Y зависимы. Теперь симметризация производится без разрыва выборочных значений X и Y с одинаковыми номерами:

$$U_3=\frac{1}{N-1}\sum_{i=1}^{N}\left[(x_i+y_j)-\frac{1}{N}\sum_{j=1}^{N}(x_j+y_j)\right]^2.$$

Пример 7. Оценивание двумерной функции распределения $F(x, y)$. Ядром минимального объема будет, очевидно,

$$f(x_1, y_1)=C(x-x_1)\,C(y-y_1),$$

где $C(t)=1$ при $t\geqslant 0$, $C(t)=0$ при $t<0$.

а. X и Y независимы. Производя сначала симметризацию f по x, а затем по y, получаем:

$$U_H=\left[\frac{1}{N}\sum_{i=1}^{N}C(x-x_i)\right]\cdot\left[\frac{1}{N}\sum_{j=1}^{N}C(y-y_j)\right].$$

б. X и Y зависимы. В этом случае симметризация производится без разрыва эмпирических пар (x_i, y_i):

$$U_3=\frac{1}{N}\sum_{i=1}^{N}[C(x-x_i)\,C(y-y_i)].$$

Полученные оценки функции распределения легко обобщаются на случай произвольной размерности k:

$$U_H = \prod_{l=1}^{k} \left[\frac{1}{N} \sum_{i=1}^{N} C(x)^{(l)} - x_i^{(l)} \right],$$

$$U_З = \frac{1}{N} \sum_{i=1}^{N} \prod_{l=1}^{k} C(x^{(l)} - x_i^{(l)}).$$

Закончив на этом с примерами, сделаем несколько замечаний, касающихся практического построения и использования U-статистик.

Как видно из приведенных выше примеров, главная трудность построения U-статистик заключается не в том, чтобы по заданной функции $\varphi(x)$ найти соответствующее ядро f, а в том, чтобы представить интересующий нас параметр в виде математического ожидания известной функции $\varphi(x)$. (Особенно наглядны в этом отношении примеры 2 и 4). После этого построение U-статистики фактически сводится лишь к соответствующей симметризации ядра, сначала по выборке объема степени параметра, затем — по всему объему выборки.

Качества U-статистик гарантируются теоремами 1—5, и эти качества весьма высоки. Следует, однако, иметь в виду, что мы строим U-статистики, не з н а я истинных распределений, тогда как свойства U-статистик зависят от этих распределений. Например, U-статистики различны для зависимых и независимых выборок (см. примеры 5—7), а на практике иногда неизвестно, имеем мы дело с наличием или отсутствием зависимости*. Другая трудность может встретиться в тех случаях, когда соответствующее математическое ожидание не существует, но мы этого не знаем, так как не знаем распределения. Например, мы хотим оценить среднее значение и берем для этого выборочное среднее (пример 1). Если же распределение окажется таким, что соответствующий интеграл расходится (как, например, для распределения Коши), то U-статистика теряет свои качества. Так, при распределении Коши выборочное среднее распределено так же, как и единичное выборочное значение, т. е. увеличение объема выборки не приводит к улучшению оценки. Поэтому в тех случаях, когда есть основания подозревать, что мы можем столкнуться с такой ситуацией, имеет смысл произвести проверку поведения U-статистики при возрастании N или на нескольких независимо полученных выборках.

---

* Заметим, что такую ситуацию можно обратить и на пользу. Например, разность U-статистик в примере 5 обычно используется как выборочный коэффициент корреляции; разность $U_З - U_Н$ в примере 7 может служить основой для построения тестов, обнаруживающих наличие зависимости и т. д.

§ 8.4. УПРОЩЕННЫЙ СПОСОБ ПОСТРОЕНИЯ U-СТАТИСТИК И РАСПРОСТРАНЕНИЕ ЭТОГО СПОСОБА НА СЛУЧАЙ ЗАВИСИМОЙ ВЫБОРКИ

Процедура построения U-статистики не всегда оказывается элементарной: необходимо отыскание ядра, получение симметричного ядра и последующая симметризация по всей выборке (см. § 8.3). Можно предложить (Симахин, Тарасенко, Шуленин [1]) несколько иной подход к оцениванию линейных функционалов, при котором не только эти сложности исчезают, но в тех случаях, когда U-статистики могут быть построены, они совпадают с оценками, вводимыми ниже, и, что более важно, этот подход дает несмещенные и состоятельные оценки и в случае зависимой выборки $\{x_i\}$, для которой теория U-статистик не развита.

Идея излагаемого подхода весьма проста и состоит в том, чтобы получать оценку линейного функционала

$$\theta = J[F(x)] = \int \varphi(x^{(1)}, \ldots, x^{(k)})\, dF(x^{(1)}, \ldots, x^{(k)}), \qquad (8.4.1)$$

подставляя в (8.4.1) вместо F его несмещенную и состоятельную оценку $F_N(x^{(1)}, \ldots, x^{(k)})$. С помощью известных теорем анализа легко доказать следующее утверждение.

Теорема 8.4.1. Пусть $F_N(x^{(1)}, \ldots, x^{(k)})$ — несмещенная и состоятельная оценка распределения k-го порядка, полученная по выборке $x_1, \ldots, x_N$. Тогда

$$J_N = \int \varphi(x^{(1)}, \ldots, x^{(k)})\, dF_N(x^{(1)}, \ldots, x^{(k)}) \qquad (8.4.2)$$

является несмещенной и состоятельной оценкой J (8.4.1).

Оценки F_N, обладающие всеми необходимыми свойствами, рассмотрены в § 7.10. Поскольку dF_N является суммой δ-функций, построение оценок J_N-чрезвычайно просто. Легко заметить, что, если $\{x_i\}$ — независимая выборка, то оценки (8.4.2) совпадают с U-статистиками. Для иллюстрации данного подхода рассмотрим несколько примеров.

Пример 1. Пусть $J = [\int x dF]^2$, выборка независима. Очевидно, $J = \iint x^{(1)} x^{(2)}\, dF(x^{(1)}, x^{(2)})$. Подставив сюда F_N из (7.10.1), сразу получаем: $J_N = [N(N-1)]^{-1} \sum\limits_{i \neq j} x_i x_j$ (Сравни с примером 2 § 8.3).

Пример 2. Пусть единственное отличие с предыдущим примером состоит в зависимости между $\{x_i\}$. U-статистики в этом случае не существует; наш подход дает $J_N^* = (N-t)^{-1} \times \sum\limits_{i=1}^{N-t} x_i x_{i+t}, t = t_2 - t_1 < m$. (Естественно, для разных t будут получаться разные оценки).

Данным способом легко получаются несмещенные и состоятельные оценки несимметричных функционалов; несимметричность может быть вызвана как самой формой функционала, так и зависимостью между выборочными значениями.

Пример 3. Пусть $J=\int f(t)\varphi(x(t))dF(x(t))$. Использование (7.10.1) и (7.10.2) дает $J_N=J_N^*=\frac{1}{N}\sum_{i=1}^{N} f(i)\varphi(x_i)$. Хотя вид этой оценки одинаков для зависимых и независимых выборок, ее свойства, конечно, будут разными.

Пример 4. Пусть $J=\int x^{(1)}\varphi(x^{(2)})dF(x^{(1)}, x^{(2)})$, и выборка зависима. Имеем: $J_N^*=(N-t)^{-1}\Sigma x_i\varphi(x_{i+t})$. При независимой выборке $J_N=[N(N-1)]^{-1}\sum_{i\neq j} x_i\varphi(x_j)$.

§ 8.5. НЕПАРАМЕТРИЧЕСКОЕ ОЦЕНИВАНИЕ НЕЛИНЕЙНЫХ ФУНКЦИОНАЛОВ ПРЯМЫМ МЕТОДОМ

Многие важные характеристики случайных величин и статистических процедур выражаются нелинейными функционалами. Например, энтропия случайной величины X, имеющей плотность распределения $g(x)$, равна

$$H(x)=-\int g(x)\ln g(x)dx;$$

фишеровская информация, играющая существенную роль для характеристики погрешности точечных оценок параметров, имеет вид:

$$I=\int [g'(x)]^2 g^{-1}(x)dx;$$

мощность ряда ранговых тестов выражается через функционал

$$J=\int g^2(x)dx;$$

и этот список можно продолжить. В ряде случаев, когда знание величины того или иного функционала оказывается необходимым, приходится прибегать к его статистическому оцениванию. Если мы имеем дело с распределением известного типа, эта задача сводится к задаче оценки параметров распределения и последующей подстановки этих оценок в функционал или прямо в конечную формулу, если соответствующий интеграл сведется к «берущемуся». В непараметрическом же случае этот метод неприменим, и необходимо развить другие подходы к оцениванию функционалов. Простейший, напрашивающийся сам собой, подход состоит в следующем

Пусть нам необходимо оценить функционал

$$J=\int_{-\infty}^{\infty}\Phi\left[g(x),\, g'(x),\ldots,\, g^{(r)}(x)\right]dx, \tag{8.5.1}$$

где g и $\{g^{(i)}\}$ — неизвестная плотность и ее производные, а $\Phi(\,.\,)$ — некоторая нелинейная функция своих аргументов. Возьмем в качестве оценки этого функционала величину J_N которая получается в результате подстановки в (8.5.1) вместо g и $\{g^{(i)}\}$ непараметрических оценок плотности и ее производных:

$$J_N=\int_{-\infty}^{\infty}\Phi\left[g_N(x),\, g_N'(x),\ldots,\, g_N^{(r)}(x)\right]dx \tag{8.5.2}$$

Так как существует несколько различных оценок плотности (см. гл. VII), возможно получение разных оценок функционала с различными, естественно, свойствами. Возникает интересная задача выяснения относительных достоинств и недостатков оценок различных типов. Для краткости все такие оценки функционалов будем называть «прямыми», имея в виду, что все они получаются благодаря непосредственной подстановке g_N в J, а сам метод получения оценок (8.5.2) — «прямым методом».

8.5.1. ПРЯМЫЕ ОЦЕНКИ НЕЛИНЕЙНЫХ ФУНКЦИОНАЛОВ, ОСНОВАННЫЕ НА ОЦЕНКЕ ПЛОТНОСТИ РОЗЕНБЛАТТА-ПАРЗЕНА

Наиболее интенсивно исследовались прямые оценки, которые получаются подстановкой в (8.5.1) оценок Розенблатта-Парзена (см. § 7.6) вида

$$g_N(x)=\frac{1}{Nh_N}\sum_{i=1}^{N}K\left(\frac{x-x_i}{h_N}\right). \tag{8.5.3}$$

Рассмотрением получающегося таким образом класса прямых оценок занимались Бхаттачарья [1], Надарая [1], Бхаттачарья и Рушас [1], Фрелик и Скотт [1] и др. Внимание, уделенное исследователями именно этому классу оценок нелинейных функционалов, объясняется двумя моментами. Во-первых, по оценке плотности $g_N(x)$ вида (8.5.3) легко получить оценку ее r-й производной (Бхаттачарья [1]):

$$g_N^{(r)}(x)=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{h_N^{r+1}}K^{(r)}\left(\frac{x-x_i}{h_N}\right). \tag{8.5.4}$$

Эти оценки производных оказываются весьма хорошими (Шустер [1]):

Лемма 8.5.1. Если $g(x)$ и ее первые $r+1$ производные ограничены и если $\{\varepsilon_N\}$ — последовательность таких положительных чисел, что $h_N=0(\varepsilon_N)$, то существуют константы $C_1>0$ и $C_2>0$ такие, что при достаточно больших N

$$P_r\{\sup_x|g_N^{(r)}(x)-g^{(r)}(x)|>\varepsilon_N\}\leqslant C_1\exp\{-G_2N\varepsilon_N^2h_N^{2r+2}\}. \quad (8.5.5)$$

Второй момент, привлекающий внимание к данным прямым оценкам функционала, связан с известной теоремой функционального анализа, утверждающей, что если некоторая последовательность функций Φ_N равномерно сходится к функции Φ, то последовательность чисел $J_N=\int_a^b\Phi_N\,dx$ сходится к величине $J=\int_a^b\Phi dx$. Тем самым в силу свойств оценок плотности Розенблатта-Парзена и ее производных гарантируется асимптотическая несмещенность и состоятельность этого класса прямых оценок.

Представляет интерес выяснение условий, при которых J_N сходится к J с вероятностью единица. Оказывается, что эти условия различны для разных функционалов. Покажем это на примерах трех функционалов, приведенных в начале параграфа; эти примеры представляют интерес и сами по себе, в силу теоретического и практического смысла функционалов, а кроме того, на этих примерах мы проиллюстрируем методы оценивания скорости сходимости оценок конкретных функционалов.

Пример 1. Оценивание интеграла от квадрата плотности,

$$J=\int_{-\infty}^{\infty}g^2(x)\,dx. \quad (8.5.6)$$

Этой задачей занимались Бхаттачарья и Рушас [1], которые показали, что прямая оценка J_N при определенных условиях является асимптотически несмещенной и состоятельной. Надарая [2] доказал, что полученная величина

$$W_N^2=\int_{-\infty}^{\infty}|g_N(x)-g(x)|^2\,dx$$

сходится к нулю с вероятностью единица. Мы здесь дадим простое доказательство сходимости J_N к J с вероятностью 1 (Дмитриев и Тарасенко [1]):

Теорема 8.5.1. Если $g(x)$ — ограниченная функция, то при $h_N=N^{-\frac{1}{4}}$, $\varepsilon_N=N^{-\frac{1}{5}}$, J_N сходится к J с вероятностью единица.

Доказательство.

$$|J_N-J|\leqslant \int |g_N^2(x)-g^2(x)|\,dx\leqslant 2\sup_x |g_N(x)-g(x)|.$$

$$Pr\{|J_N-J|>\varepsilon_N\}\leqslant P_r\left\{\sup_x |g_N(x)-g(x)|>\frac{\varepsilon_N}{2}\right\}.$$

Применяя лемму 8.5.1, получаем:

$$Pr\{|J_N-J|>\varepsilon_N\}\leqslant C_1\exp\left\{-\frac{C_2}{4}N^{\frac{1}{10}}\right\}, \tag{8.5.7}$$

а так как ряд $\sum_{N=1}^{\infty}\exp(-CN^{\frac{1}{10}})<\infty$, то, согласно лемме Бореля-Кантелли, $J_N\to J$ с вероятностью единица.

Замечание. Условие $h_N=N^{-\frac{1}{4}}$ характеризует максимальную допустимую скорость убывания h_N. При меньших скоростях убывания, например, $h_N=N^{-\frac{1}{4+k}}$, $k\geqslant 0$, теорема остается верной, если ε_N таково, чтобы $h_N=0(\varepsilon_N)$ (см (8.5.5)).

Пример 2. Оценивание фишеровской информации*

$$J=\int_{-\infty}^{\infty}\frac{[g^{(1)}(x)]^2}{g(x)}\,dx \tag{8.5.8}$$

величиной

$$J_N=\int_{-K_N}^{K_N}\frac{[g_N^{(1)}(x)]^2}{g_N(x)}\,dx. \tag{8.5.9}$$

Функционал (8.5.8) отличается от рассмотренного в примере 1 не только наличием производной, но и тем, что $g(x)$ функция с неограниченно убывающими хвостами стоит в знаменателе подынтегрального выражения. При конкретизации характера сходимости J_N к J это приводит к появлению границ $\pm K_N$ в (8.5.9) и необходимости учета скорости спадания хвостов $g(x)$. Пусть, для определенности, эту скорость можно оценить монотонной и строго возрастающей функцией $H(y)$, такой, что для всех y имеет место

$$\sup_{|x|<y}\frac{1}{g(x)}\leqslant H(y). \tag{8.5.10}$$

Тогда верна следующая

* Эту задачу рассматривал и Бхаттачарья ([1], теорема 3, неравенство (7) § 4). Однако им допущена ошибка, исправляемая нижеследующей теоремой 8.5.2 (Дмитриев и Тарасенко [1]).

Теорема 8.5.2. Если $g(x)$ и $g'(x)$ ограниченные функции, то при выполнении (8.5.10) и $h_N=N^{-\frac{1}{6}}$ и $K_N=H^{-1}(N^{\frac{1}{28}})$, J_ν (8.5.9) сходится к J (8.5.8) с вероятностью единица.

Доказательство. Представим разность J_N-J в виде суммы $U_N+V_N+e_N$, где

$$U_N=\int\limits_{|x|\leqslant K_N}\left[\frac{g'_N(x)}{g_N(x)}\right]^2\left[\frac{g_N(x)-g(x)}{g(x)}\right]g(x)\,dx,$$

$$V_N=\int\limits_{|x|\leqslant K_N}\left\{\left[\frac{g'_N(x)}{g_N(x)}\right]^2-\left[\frac{g'(x)}{g(x)}\right]^2\right\}g(x)\,dx,$$

$$e_N=-\int\limits_{|x|>K_N}\frac{[g'(x)]^2}{g(x)}\,dx.$$

Если

$$\sup_{|x|\leqslant K_N}|g_N(x)-g(x)|\leqslant\varepsilon_N \text{ и } \sup_{|x|\leqslant K_N}|g'_N(x)-g'(x)|\leqslant\varepsilon_N, \quad (8.5.11)$$

то в силу ограниченности g и g' существуют такие константы C и B, что при достаточно больших N имеют место следующие неравенства:

$$\sup_{|x|\leqslant K_N}\left|\frac{g_N(x)-g(x)}{g(x)}\right|\leqslant\varepsilon_N H(K_N),$$

$$\sup_{|x|\leqslant K_N}\left|\frac{g'_N(x)}{g_N(x)}\right|^2\leqslant CH^2(K_N),$$

$$\sup_{|x|\leqslant K_N}\left|\left[\frac{g'_N(x)}{g_N(x)}\right]^2-\left[\frac{g'(x)}{g(x)}\right]^2\right|\leqslant 2BC\varepsilon_N H^4(K_N),$$

где $|g(x)|\leqslant B<\infty$, $|g'(x)|^2\leqslant C<\infty$. Из этих неравенств следует, что

$$|U_N|\leqslant C\varepsilon_N H^3(K_N) \text{ и } |V_N|\leqslant 2BC\varepsilon_N H^4(K_N). \quad (8.5.12)$$

В самом деле, зададимся $\varepsilon>0$. В предположении ограниченности J при достаточно больших N и произвольном $\varepsilon>0$, $|e_N|<\frac{\varepsilon}{2}$. Если выбрать ε_N так, что $\varepsilon_N=\varepsilon[4BC^2H^4(K_N)]^{-1}=$ $=\varepsilon/4BCN^{\frac{1}{7}}$, то при $N\to\infty$ из (8.5.11) следует (8.5.12). Отсюда,

$$Pr\{|J_N-J|>\varepsilon\}<Pr\{\sup_{|x|\leqslant K_N}|g_N(x)-g(x)|\geqslant\varepsilon/4BCN^{\frac{1}{7}}\}+$$

$$+Pr\{\sup_{|x|\leqslant K_N}|g'_N(x)-g'(x)|\geqslant\varepsilon/4BCN^{\frac{1}{7}}\}<$$

$$<2Pr\{\sup_{|x|\leqslant K_N}|g'_N(x)-g'(x)|\geqslant\varepsilon/4BCN^{\frac{1}{7}}\}\leqslant$$

$$\leqslant 2Pr\{\sup_{-\infty\leqslant x\leqslant\infty}|g'_N(x)-g'(x)|\geqslant\varepsilon/4BCN^{\frac{1}{7}}\}. \qquad (8.5.13)$$

Выбирая $h_N=0(\varepsilon_N)$ и применяя лемму 8.5.1, получим верхние границы для вероятности в правой части (8.5.13):

$$Pr\{|J_N-J|>\varepsilon\}<2C_1\exp\left\{-\frac{C_2\varepsilon^2}{(4BC)^2}N^{\frac{1}{21}}\right\}, \qquad (8.5.14)$$

а так как $\sum_1^\infty \exp(-AN^{\frac{1}{21}})<\infty$, то из леммы Бореля-Кантелли следует истинность теоремы 8.5.2.

Пример 3. Оценивание дифференциальной энтропии

$$J=-\int_{-\infty}^{\infty} g(x)\ln g(x)\,dx \qquad (8.5.15)$$

величиной

$$J_N=-\int_{-K_N}^{K_N} g_N(x)\ln g_N(x)dx. \qquad (8.5.16)$$

Теорема 8.5.3. Если $g(x)$ ограничена сверху и выполняется неравенство (8.5.10), то при $h_N=N^{-\frac{1}{4}}$ и $K_N=H^{-1}(N^{\frac{1}{10}})$ оценка J_N (8.5.16) сходится к J (8.5.15) с вероятностью единица.

Доказательство. Представим J_N-J в виде

$$J_N-J=U_N^*+V_N^*+e_N^*,$$

где

$$U_N^*=\int_{|x|\leqslant K_N} g(x)\ln\frac{g_N(x)}{g(x)}\,dx,$$

$$V_N^*=\int_{|x|\leqslant K_N}[g_N(x)-g(x)]\ln g_N(x)\,dx,$$

$$e_N^*=-\int_{|x|\geqslant K_N} g(x)\ln g(x)\,dx.$$

Используя неравенства

$$|\ln y|\leqslant\left|\frac{1}{y}-1\right|+|y-1|,\ y\geqslant 0,$$

$$|\ln y| < \frac{1}{y} + y, \quad y \geqslant 0,$$

можем записать:

$$|U_N^*| \leqslant \int\limits_{|x| \leqslant K_N} \left[\left| \frac{g(x) - g_N(x)}{g_N(x)} \right| + \left| \frac{g_N(x) - g(x)}{g(x)} \right| \right] g(x)\,dx,$$

$$|V_N^*| < \int\limits_{|x| \leqslant K_N} |g_N(x) - g(x)| \left[\frac{g(x)}{g_N(x) - g(x)} - g_N(x) \right] dx.$$

Если

$$\sup_{|x| \leqslant K_N} |g_N(x) - g(x)| \leqslant \varepsilon_N \tag{8.5.17}$$

и выполнено условие (8.5.10), то при достаточно больших N

$$|U_N^*| \leqslant 2H(K_N)\varepsilon_N, \quad |V_N^*| < [H^2(K_N) + 1]\,\varepsilon_N. \tag{8.5.18}$$

Так как $|J| < \infty$, то при больших N и $\varepsilon > 0$ $|e_N'| < \varepsilon/2$. Выберем $\varepsilon_N = \varepsilon/2H^2(K_N) = \varepsilon/2N^{\frac{1}{5}}$. Тогда из (8.5.17) следует (8.5.18), в силу чего

$$Pr\{|J_N - J| > \varepsilon\} < Pr\{\sup_{|x| < K_N} |g_N(x) - g(x)| \geqslant \varepsilon/8N^{\frac{1}{5}}\} \leqslant$$

$$\leqslant Pr\{\sup_{-\infty < x < \infty} |g_N(x) - g(x)| \geqslant \varepsilon/8N^{\frac{1}{5}}\}.$$

Так как $h_N = 0(\varepsilon_N)$, то, согласно лемме 8.5.1, имеем

$$Pr\{|J_N - J| > \varepsilon\} < C_1 \exp\left\{ -\frac{C_2}{4} \varepsilon^2 N^{\frac{1}{10}} \right\}, \tag{8.5.19}$$

и так как $\sum\limits_1^\infty \exp\{-CN^{\frac{1}{10}}\} < \infty$, то теорема доказана.

Теперь можно сделать некоторые общие выводы о прямых оценках нелинейных функционалов при использовании оценок плотностей Розенблатта-Парзена.

Во-первых, равномерная по x сходимость g_N к g гарантирует асимптотическую несмещенность и состоятельность (т. е. сходимость по вероятности) оценок (8.5.2).

Во-вторых, выбирая подходящим для каждого функционала образом параметр h_N оценки g_N и (в случае необходимости) параметр K_N оценки J_N, можно обеспечить сходимость J_N к J с вероятностью единица.

Рассмотрим теперь некоторые вопросы, связанные со сходимостью оценок (8.5.2) по вероятности.

В § 7.6 вычислены среднее и дисперсия для $g_N(x)$. Эти

результаты легко обобщаются на случай произвольного целого $r \geqslant 0$ в (8.5.4). Опуская простые выкладки, имеем:

$$\lim_{N\to\infty} Eg_N^{(r)}(x) = g^{(r)}(x); \tag{8.5.20}$$

$$Dg_N^{(r)}(x) = \frac{1}{Nh_N^{2r+1}} \int_{-\infty}^{\infty} [K^{(r)}(z)]^2 g(x - zh_N)\, dz + \\ - \frac{1}{Nh_N^{2r}} \left[\int_{-\infty}^{\infty} K^{(r)}(z)\, g(x - zh_N)\, dz \right]^2. \tag{8.5.21}$$

Из (8.5.21) следует, что при $N \to \infty$, $h_N \to 0$ и $Nh_N^{2r+1} \to \infty$

$$\lim_{N\to\infty} Nh_N^{2r+1} Dg_N^{(r)}(x) = g(x) \int_{-\infty}^{\infty} [K^{(r)}(z)]^2\, dz. \tag{8.5.22}$$

Лемма 8.5.3. Если g и ее первые $r+1$ производные ограничены, и $\{\varepsilon_N\}$ — такая последовательность положительных чисел, что $h_N = 0(\varepsilon_N)$, то существует такая положительная константа $C < \infty$, что при достаточно больших N

$$Pr\{\sup_x |g_N^{(r)}(x) - g^{(r)}(x)| > \varepsilon_N\} < \frac{C}{Nh_N^{2r+1}\varepsilon_N^2}. \tag{8.5.23}$$

Доказательство. Как показал Бхаттачарья [1], справедливо следующее соотношение:

$$\sup_x \left| \frac{1}{h_N^{r+1}} \int_{-\infty}^{\infty} K^{(r)}\left(\frac{x-y}{h_N}\right) g(y)\, dy - g^{(r)}(x) \right| \leqslant \bar{C} h_N,$$

где $\bar{C} = \sup_x |g^{(r+1)}(x)| \int |u| K(u)\, du$. Запишем очевидное неравенство

$$|g_N^{(r)}(x) - g^{(r)}(x)| < |g_N^{(r)}(x) - Eg_N^{(r)}(x)| + \bar{C} h_N.$$

Так как $h_N = 0(\varepsilon_N)$, то для достаточно больших N можно положить $\bar{C} h_N = \varepsilon_N/2$; тогда

$$Pr\{|g_N^{(r)}(x) - g^{(r)}(x)| > \varepsilon_N\} \leqslant Pr\left\{|g_N^{(r)}(x) - Eg_N^{(r)}(x)| > \frac{\varepsilon_N}{2}\right\} \leqslant \\ \leqslant \frac{4Dg_N^{(r)}(x)}{\varepsilon_N} \leqslant \frac{4 \sup_x Dg_N^{(r)}(x)}{\varepsilon_N^2}. \tag{8.5.24}$$

В силу (8.5.24), с учетом (8.5.21) имеем:

$$Pr\{\sup_x |g_N^{(r)}(x) - g^{(r)}(x)| > \varepsilon_N\} \leqslant \frac{4 \sup_x Dg_N^{(r)}(x)}{\varepsilon_N^2} =$$

$$= \frac{4\sup_x \left\{ \int_{-\infty}^{\infty} [K^{(r)}(z)]^2 g(x-h_N z)dz - h_N \left[\int_{-\infty}^{\infty} K^{(r)}(z) g(x-zh_N)dz \right]^2 \right\}}{Nh_N^{2r+1}\varepsilon_N^2} \leqslant$$

$$\leqslant \frac{C}{Nh_N^{2r+1}\varepsilon_N^2}, \qquad (8.5.25)$$

где $C = 4\sup_x g(x) \int_{-\infty}^{\infty} K^{(r)}(z)]^2 dz$. Лемма доказана.

Так как ряд $\Sigma N^{-\alpha}$ при любых $\alpha \leqslant 1$ расходится, то граница (8.5.23) гарантирует лишь сходимость оценок по вероятности, тогда как (8.5.5) при правильном выборе ε_N и h_N обеспечивает сходимость с вероятностью единица. Вопрос о практическом сопоставлении оценок, обладающих разными типами сходимости, интересен сам по себе и будет рассмотрен в другом параграфе. Здесь же дадим теорему, характеризующую скорость сходимости $g_N^{(r)}(x)$ к $g^{(r)}(x)$.

Теорема 8.5.4. Если $g(x)$ и ее первые $r+1$ производные ограничены, и если последовательности $\{h_N\}$ и $\{b_N\}$ таковы, что при $N \to \infty$ $h_N b_N = 0(1)$ и $Nh_N^{2r+1}/b_N^2 \to \infty$, то величина $[b_N \sup_x |g_N^{(r)} g^{(r)}|]$ сходится при $N \to \infty$ к нулю по вероятности.

Доказательство. Следует непосредственно из (8.5.25), если положить $b_N = \varepsilon/\varepsilon_N$, где $\varepsilon > 0$ — сколь угодно малая постоянная величина.

Следствие. Если $g(x)$ и ее первые $r+1$ производные ограничены, $h_N = N^{-1/(2r+2)}$ и $0 < \theta < 1/2(2r+2)$, то величина $[N^{\theta} \sup_x |g_N^{(r)}(x) - g^{(r)}(x)|]$ при $N \to \infty$ сходится к нулю по вероятности.

8.5.2. ПРЯМЫЕ ОЦЕНКИ НЕЛИНЕЙНЫХ ФУНКЦИОНАЛОВ, ОСНОВАННЫЕ НА ГИСТОГРАММЕ

Учитывая широкое практическое употребление гистограммы в качестве оценки плотности, интересно рассмотреть, что даст использование гистограммы для получения прямых оценок функционалов.

Первое ограничение, накладываемое особенностями гистограммы, состоит в том, что функционалы, содержащие производные плотности, не могут быть оценены прямым методом: производная гистограммы является суммой δ-функций. Поэтому будем рассматривать только случай, когда (8.5.1) не содержит производных плотности.

Представим гистограмму (см. § 7.4) аналитически формулой

$$g_N(x)=\sum_k\left[\frac{\pi(x,\Delta_k)}{\Delta_k}\cdot\sum_{i=1}^{N}\frac{\pi(x_i,\Delta_k)}{N}\right], \qquad (8.5.26)$$

где $x_1, \ldots, x_N$ — выборка объема N, Δ_k — величина k-го интервала группировки,

$$\pi(x,\Delta_k)=\begin{cases}1, & x\in\Delta_k,\\ 0, & x\in\Delta_k.\end{cases}$$

Подстановка (8.5.26) в (8.5.1) дает следующую прямую оценку:

$$J_N=\sum_k\Phi\left[\frac{\tilde p_k}{\Delta_k}\right]\cdot\tilde p_k, \qquad (8.5.27)$$

где $\tilde p_k=N^{-1}\sum_{i=1}^{N}\pi(x_i,\Delta_k)$ есть относительная частота события $X\in\Delta_k$.

Рассмотрим, какими свойствами обладает прямая оценка (8.5.27). Известно, что относительная частота $\tilde p_k$ имеет асимптотически нормальное распределение со средним p_k и дисперсией $\sigma^2_{kN}=p_k(1-p_k)/N$, где p_k — истинное значение вероятности события $X\in\Delta_k$. Отсюда, последовательность чисел $\xi_{kN}=\tilde p_k-p_k$ сходится к нулю по вероятности. Если теперь ограничиться классом функций $\Phi(t)$, разложимых в ряд Тэйлора для всех t, то можно записать

$$J_N=\sum_k\sum_{m=0}^{\infty}\frac{1}{m!}\cdot\Phi^{(m)}\left(\frac{p_k}{\Delta_k}\right)\left(\frac{\xi_{kN}}{\Delta_k}\right)^m(p_k+\xi_{kN}), \qquad (8.5.28)$$

а при достаточно больших N можно ограничиться учетом лишь первых трех членов этого ряда:

$$J_N\doteq\sum_k[\alpha_k+\theta_k\xi_{kN}+\beta_k\xi^2_{kN}], \qquad (8.5.29)$$

где

$$\alpha_k=p_k\Phi\left(\frac{p_k}{\Delta_k}\right);\quad \theta_k=\Phi\left(\frac{p_k}{\Delta_k}\right)+\frac{p_k}{\Delta_k}\Phi'\left(\frac{p_k}{\Delta_k}\right);$$

$$\beta_k=\Phi'\left(\frac{p_k}{\Delta_k}\right)\cdot\frac{1}{\Delta_k^2}+\frac{p_k}{2}\cdot\Phi''\left(\frac{p_k}{\Delta_k}\right).$$

Усредняя последнее равенство, имеем:

$$EJ_N\doteq\sum_k p_k\Phi\left(\frac{p_k}{\Delta_k}\right)+\sum_k\beta_k\sigma^2_{kN}\xrightarrow[N\to\infty]{}\sum_k p_k\Phi\left(\frac{p_k}{\Delta_k}\right). \qquad (8.5.30)$$

Вычисление дисперсии (с учетом того, что $E\xi_{kN}\xi_{tN} = -p_k p_t/N$), приводит к результату

$$\lim_{N\to\infty} NDJ_N = \sum_k \theta_k^2 p_k(1-p_k) - \sum_{k\neq t}\theta_k\theta_t p_k p_t < \infty.$$

Итак прямые оценки функционалов, получаемые с помощью гистограммы, хотя и имеют убывающую с ростом N дисперсию, обладают смещением, которое зависит от $g(x)$ и, следовательно, в непараметрическом случае неизвестно. Конечно, можно уменьшить смещенность оценки (8.5.27) путем увеличения числа интервалов группировки, но тогда потребуется соответствующее увеличение объема выборки. Эта особенность данного типа прямых оценок делает их малоинтересными для оценивания. Однако существует область статистических приложений, в которой данные оценки могут с успехом применяться. Речь идет о задаче согласия (см. гл. IX). Из (8.5.30) следует, что (если Ф выбрана удачно) среднее значение J_N будет сходиться к разным величинам при разных распределениях. Этот факт и можно использовать, так как нулевая гипотеза F в задаче согласия известна, и остается лишь установить отличие $J_N(G)$ от $J_N(F)$. Известным примером является χ^2-критерий, но очевидно, что это — не единственный возможный тест согласия, основанный на гистограмме.

8.5.3. ПРЯМЫЕ ОЦЕНКИ НА ПОЛИГРАММЕ

Полиграмма является непараметрической оценкой плотности (см. § 7.5) и также может быть использована при оценивании функционалов. По тем же причинам, что и гистограмма, полиграмма не пригодна в тех случаях, когда функционал содержит производные плотности. Рассмотрим свойства прямых оценок, получаемых с помощью полиграммы, на примере использования полиграммы первого порядка.

Пусть функционал имеет вид

$$J = \int F[g(x)]g(x)\,dx. \tag{8.5.31}$$

Подставляя сюда вместо $g(x)$ полиграмму $g_N(x)$, имеем

$$J_N = \int F\left[\frac{1}{N+1}\sum_{i=1}^{N+1}\frac{\pi(x,\Delta_i)}{\Delta_i}\right]\frac{1}{N+1}\cdot\sum_{i=1}^{N+1}\frac{\pi(x,\Delta_i)}{\Delta_i}\,dx =$$
$$= \frac{1}{N+1}\sum_{i=1}^{N+1} F\left[\frac{1}{(N+1)\Delta_i}\right]. \tag{8.5.32}$$

Если Δ_1 и/или Δ_{N+1} бесконечны, то соответствующие члены в (8.5.32) должны приниматься равными нулю.

Свойства статистики (8.5.32) существенно определяются не только распределением выборки, но и видом функции F Можно сделать лишь следующие общие утверждения: 1. Если случайные величины $F\left[\frac{1}{(N+1)\Delta_i}\right]$ имеют первые два момента, то в силу асимптотической независимости выборочных интервалов и центральной предельной теоремы, оценка J является асимптотически нормальной. 2. Если функция F такова, что моменты величин $F\left[\frac{1}{(N+1)\Delta_i}\right]$ неограничены, то J_N также не имеет моментов. Суждения о свойствах прямых оценок на полиграмме можно несколько конкретизировать, если воспользоваться представлением выборочных интервалов через экспоненциально распределенные величины (см. § 4.4). Из (4.4.17) следует, что

$$\frac{1}{(N+1)\Delta_i} = \frac{1-\frac{i}{N+1}}{Y_i} \cdot h\left[G^{-1}(A_i)\right], \qquad (8.5.33)$$

где $h(x)=g(x)/[1-G(x)]$ — функция интенсивности. При вычислении асимптотики первого момента J_N можно воспользоваться сходимостью порядковых статистик к квантилям, что дает

$$\frac{1}{(N+1)\Delta_i} \to \frac{g\left[G^{-1}\left(\frac{i}{N+1}\right)\right]}{Y_i}. \qquad (8.5.34)$$

Таким образом,

$$\lim_{N\to\infty} J_N = \lim_{N\to\infty} \frac{1}{N+1} \sum_{i=1}^{N+1} \int F\left[\frac{g\left(G^{-1}\left(\frac{i}{N+1}\right)\right)}{Y_i}\right] e^{-Y_i} d\ Y_i \qquad (8.5.35)$$

и дальнейшие рассуждения можно провести, лишь конкретизируя вид функции F.

Предположим, например, что оценивается дифференциальная энтропия распределения $g(x)$. Тогда $F(g)=-\ln g$, и мы имеем:

$$\lim_{N\to\infty} J_N = \lim_{N\to\infty} \frac{-1}{N+1} \sum_{i=1}^{N+1} \ln g\left[G^{-1}\left(\frac{i}{1+1}\right)\right] + \int_{-\infty}^{\infty} \frac{1}{y} e^{-y}\, dy =$$
$$= -\int g(x) \ln g(x)\, dx - C = J - C,$$

где C — константа Эйлера. Таким образом, прямая оценка функционала энтропии, основанная на выборочных интерва-

лах, имеет асимптотическое смещение, равное $-C$. Легко показать, что дисперсия этой оценки при $N\to\infty$ стремится к нулю, что и дает окончательное представление о ее свойствах.

В качестве примера, в котором J_N не является асимптотически нормальной, рассмотрим оценивание интеграла от квадрата плотности. В этом случае (8.5.32) приобретает вид

$$J_N = \frac{1}{(N+1)^2}\sum \Delta_i^{-1}, \qquad (8.5.36)$$

и J_N не имеет конечных моментов. Исследование таких случаев сложно и представляет самостоятельный интерес.

§ 8.6. ОЦЕНИВАНИЕ НЕЛИНЕЙНЫХ ФУНКЦИОНАЛОВ КВАЗИ-U-СТАТИСТИКАМИ

Можно указать следующие недостатки прямых оценок нелинейных функционалов, рассмотренных в предыдущем параграфе. Во-первых, при практическом оценивании оказывается необходимым выполнение интегральных операций над непрерывными случайными функциями, что является трудоемким делом и связано с внесением дополнительных погрешностей. Во-вторых, для ряда функционалов прямые оценки оказываются весьма медленно сходящимися, что требует очень больших объемов выборки. Поэтому представляют интерес поиск и исследование других методов оценивания нелинейных функционалов. Одному из таких методов и посвящен данный параграф.

Идея состоит в том, чтобы использовать непараметрическую оценку плотности лишь до такой степени, чтобы можно было привести оцениваемый функционал к виду, допускающему его дальнейшее оценивание с помощью U-статистики.

Известно (см. § 8.3), что U-статистиками можно оценивать функционалы вида

$$J = \int f(x)g(x)dx, \qquad (8.6.1)$$

где $f(x)$ — заданная функция случайной величины X, а $g(x)$ — неизвестная плотность. Любой нелинейный функционал можно привести к виду

$$J = \int F[g(x)]g(x)dx, \qquad (8.6.2)$$

но поскольку $g(x)$, а следовательно, и $F[g(x)]$ неизвестны, построение U-статистик исключается. Однако, если в $F[g(x)]$ подставить вместо $g(x)$ ее непараметрическую оценку $g_N(x)$ полученная функция $f_N(x)=F[g_N(x)]$ оказывается известной, (8.6.2) приобретает вид (8.6.1), что и позволяет далее

применить методы, используемые при построении U-статистик. Будем называть получаемые таким способом оценки функционалов квази-U-статистиками. Можно ожидать, что квази-U-статистики будут обладать хорошими свойствами. Во-первых, какая-то доля высоких качеств U-статистик должна перейти на квази-U-статистики. Во-вторых, практическое получение оценок функционалов квази-U-статистиками не требует выполнения интегральных операций, что также является достоинством.

Прежде чем переходить к рассмотрению свойств квази-U-статистик, укажем, что существует три типа таких оценок, которые различаются способом использования выборки. Предположим, что оценивание плотности производится по выборке объема N_1, симметризация — по выборке объема N_2, а общий объем выборки $N \leqslant N_1+N_2$. Ясно, что качество оценки будет зависеть от того, перекрываются ли указанные подвыборки и каково соотношение их объемов. Для конкретности квази-U-статистики при $N_1=N_2=N$ будем называть оценками I типа, при $N_1+N_2=N$ — оценками II типа и при $N_1+N_2>N$ — оценками III типа.

По-видимому, нельзя сделать суждений о свойствах квази-U-статистик безотносительно к виду функции F в (8.6.2). Поэтому мы проведем рассмотрение этих свойств для двух достаточно широких классов функций F и для некоторых конкретных F, представляющих самостоятельный интерес.

8.6.1. КВАЗИ-U-СТАТИСТИКИ НА ОЦЕНКАХ РОЗЕНБЛАТТА—ПАРЗЕНА—БХАТТАЧАРЬЯ

Предположим, что плотность $g(x)$ оценивается по выборке $x_1, \ldots, x_{N_1}$ способом, описанным в § 7.6, а симметризация проводится по выборке $y_1, \ldots, y_{N_2}$. Тогда оценка I типа имеет вид

$$J_N = \frac{1}{N} \sum_{j=1}^{N} F\left[\frac{1}{Nh_N} \sum_{i=1}^{N} K\left(\frac{x_j - x_i}{h_N}\right)\right], \tag{8.6.3}$$

а оценки II и III типа получаются в виде:

$$J_N = \frac{1}{N_2} \sum_{j=1}^{N_2} F\left[\frac{1}{N_1 h_1} \sum_{i=1}^{N_1} K\left(\frac{y_j - x_i}{h_{N_1}}\right)\right]. \tag{8.6.4}$$

Рассмотрим сначала класс функций $F(t)$, разложимых в ряд Тэйлора во всех точках $t \in [0, \infty]$ (Серых и Тарасенко [1]). Так как $g_N(x)$ является асимптотически несмещенной и состоятельной оценкой $g(x)$ (см. § 7.6), ее можно представить в виде

$$_N(x)=g(x+\Delta g(x)=g(x)+\Delta_N(x)+\xi_N(x), \qquad (8.6.5)$$

инная плотность, Δ_N — смещение, а ξ_N — асимпто-
мальная случайная величина с нулевым средним;
ξ_N величины ξ_N и величина Δ_N стремятся к ну-
астании N, поэтому при больших N следует ожи-
ей сходимости ряда Тэйлора для $F(g)$ по степе-
и $g_N-g=\Delta g$:

$$F[g_N(x)]=F[g(x)+\Delta g(x)]=$$

$$=F[g(x)]+\sum_{m=0}^{\infty}\frac{1}{m!}F_g^{(m)}[g(x)](\Delta g). \qquad (8.6.6)$$

о кроме указанных свойств величины $\Delta_N(x)$ и
ают свойствами

$$\int\Delta_N(x)\,dx=0, \quad \int\xi_N(x)\,dx=0, \qquad (8.6.7)$$

яются простым следствием условий нормировки

Найдем смещение квази-U-статистики. Подставив (8.6.6) (8.6.3), производя усреднение и переходя к пределу, имеем:

$$\lim_{\to\infty}(EJ_N-J)=\lim_{N\to\infty}\sum_{m=1}^{\infty}\frac{1}{m!}\sum_{s=0}^{m}\binom{m}{s}\int F_g^{(m)}\Delta_N^{m-s}E\xi_N^s g(x)dx. \qquad (8.6.8)$$

ак как при $N\to\infty$ $\Delta_N\to 0$; $E\xi_N^{2k+1}=0$; $k=0, 1, 2, \ldots$; $E\xi_N^{2k}=(2m)!\,(D\xi_N)^{2k}/2^m\cdot m!$ (см. Гнеденко [3]); и $D\xi_N\to 0$, то редел правой части (8.6.8) равен нулю, откуда следует

Теорема 8.6.1. Если $F(t)$ разложима в ряд Тэйлора для сех $t\geqslant 0$, квази-U-статистика I типа является асимптотически несмещенной оценкой J.

Рассмотрим теперь асимптотическое поведение дисперсии вази-U-статистики (8.6.3). В силу сходимости последовательностей Δ_N и ξ_N к нулю при $N\to\infty$ можно ограничиться четом первых двух членов ряда в (8.6.6); тогда

$$J\doteq\frac{1}{N}\sum_{i=1}^{N}\{F[g(x_i)]+F_g^{(1)}[g(x_i)]\cdot\Delta g(x_i)+$$

$$+\frac{1}{2}F_g^{(2)}[g(x_i)][\Delta g(x_i)]^2\}=\frac{1}{N}\sum_{i=1}^{N}(F_i+A_i+B_i), \qquad (8.6.9)$$

це для краткости введены очевидные обозначения. Производя соответствующие вычисления по формуле

$$DJ_N=EJ_N^2-E^2J_N,$$

олучим:

$$DJ_N \geqslant \frac{1}{N}\{(EF^2-E^2F)+(EA^2+E^2A)+(EB^2-E^2B)+$$
$$+2[E(FA)-(EF)(EA)]+2[E(FB)-(EF)(EB]+$$
$$+2]E(AB)-(EA)(EB)]\},$$

где знак $\geqslant$ получается из-за неучтенных членов ряда Тэйлора, а также благодаря тому, что усреднение производилось без учета зависимости между величинами $\xi_N(x_i)$ и $\xi_N(x_j)$, тогда как они коррелированы в силу (8.6.7)*. Каждая из первых трех разностей в (8.6.10) положительна в силу неравенства Йенсена $\int f^2 g dx \geqslant (\int f g dx)^2$, а так как при $N\to\infty$ все разности кроме первой стремятся к нулю, то доказана

Теорема 8.6.2. Если $F(t)$ разложима в ряд Тэйлора для всех $t\geqslant 0$ и интеграл $\int F^2(g)g dx$ существует, дисперсия квази-U-статистики I типа убывает обратно пропорционально объему выборки:

$$\lim_{N\to\infty} NDJ_N = \int F^2[g(x)]\,g(x)\,dx - J^2. \qquad (8.6\,11)$$

Из теорем 8.6.1, 8.6.2, аддитивной структуры оценки (8.6 3) и центральной предельной теоремы легко следует

Теорема 8.6.3. Если $F(t)$ разложима в ряд Тэйлора во всех точках $t\geqslant 0$, и интеграл $\int F^2(g)g dx$ существует, квази-U-статистика I типа распределена асимптотически нормально

Очевидным, но важным следствием предыдущих теорем является

Теорема 8 6.4 Если $F(t)$ представима рядом Тэйлора во всех точках $t\geqslant 0$, и интеграл $\int F(g)g dx$ существует, квази-U-статистика I типа сходится к J в среднеквадратическом смысле

Приведенные выше теоремы характеризуют лишь свойства квази-U-статистик для определенного класса функционалов; само же оценивание должно производиться по формуле (8 6 3) В некоторых случаях функция F может оказаться сложной и процедура оценивания станет фактически громоздкой Поэтому имеет смысл рассмотреть вопросы приближенного вычисления квази-U-статистик за счет аппроксимации сложной функции F. Поскольку при приближениях всегда приходится накладывать ограничения на класс аппроксимируемых функций, мы снова оказываемся перед возможностью рассмотрения различных классов приближенных квази-U-ста-

* Легко показать, что коэффициент корреляции между $g_N(x_i)$ и $g_N(x_j)$, $i\neq j$ стремится к нулю при $N\to\infty$, что оправдывает дальнейшие рассуждения

тистик. Следуя работе Дмитриева и Тарасенко [1], остановимся на классе, порождаемом функциями $F(t)$, которые разложимы в степенной ряд по t:

$$F(t)=\sum_{k=1}^{\infty} a_k t^k. \tag{8.6.12}$$

Если оборвать ряд (8.6.12) на n-м члене, то приближенное значение функционала J запишется в виде:

$$\tilde{J}=\sum_{k=1}^{n} a_k \int g^{k+1}(x)\,dx, \tag{8.6.13}$$

и приближенная квази-U-статистика I типа для J будет иметь вид

$$\tilde{J}_N=\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{n}\frac{a_k}{(Nh_N)^k}\left[\sum_{j=1}^{N} K\left(\frac{x_j-x_i}{h_N}\right)\right]^k. \tag{8.6.14}$$

Таким образом, оценивание J сводится к оцениванию интегралов от степенной неизвестной плотности. Последняя задача рассматривалась в работе Дмитриева и Тарасенко [1].

Приведем теперь пример оценивания конкретного функционала. Пусть

$$J=\int_{-\infty}^{\infty} g^2(x)\,dx, \tag{8.6.15}$$

квази-U-статистика I типа для этого функционала оказывается очень простой:

$$J_N=\frac{1}{N(N-1)h_N}\sum_{i\neq j} K\left(\frac{x_i-x_j}{h_N}\right). \tag{8.6.16}$$

Среднее значение J_N равно

$$\begin{aligned} EJ_N&=\frac{1}{h_N}\iint K\left(\frac{y-x}{h_N}\right) g(x)\,g(y)\,dx\,dy= \\ &=\iint K(u)\,g(uh_N+x)\,g(x)\,du\,dx. \end{aligned} \tag{8.6.17}$$

Так как $h_N\to 0$ при $N\to\infty$, то J_N является асимптотически несмещенной оценкой:

$$\lim_{N\to\infty} EJ_N=\int g^2(x)\,dx. \tag{8.6.18}$$

Вычислим теперь дисперсию J_N, введя для краткости обозначение

$$b_{ij}=\frac{1}{h_N} K\left(\frac{x_i-x_j}{h_N}\right).$$

$$J_N^2=\frac{1}{N^2(N-1)^2}\left(\sum_{j\neq i} b_{ij}\right)^2=\frac{1}{N^2(N-1)^2}\{\sum\sum b_{ij}^2+$$
$$+\sum\sum b_{ij}b_{ji}+\sum\sum\sum\sum b_{ji}b_{pi}+\sum\sum\sum b_{ji}b_{jp}+$$
$$+\sum\sum\sum b_{ji}b_{ip}+\sum\sum\sum b_{ji}b_{ij}+\sum\sum\sum b_{ji}b_{ti}\}. \quad (8.6.19)$$

Во всех суммах индексы, характеризующие каждый член, не совпадают. Предполагая симметричность весовой функции K, имеем $b_{ji}=b_{ij}$. С учетом этого, усреднение (8.6.19) дает:

$$EJ_N^2=\frac{1}{N(N-1)}[2Eb^2+(N-2)(N-3)E^2b+$$
$$+4(N-2)E_xE_y^2b],$$

где

$$E^2b=\frac{1}{h_N^2}\iint K^2\left(\frac{x-y}{h_N}\right)g(x)g(y)\,dx\,dy,$$
$$E_xE_y^2b=\frac{1}{h_N^2}\int\left[\int K^2\left(\frac{x-y}{h_N}\right)g(y)\,dy\right]^2 g(x)dx.$$

Отсюда, дисперсия J_N равна

$$DJ_N=\frac{4}{N}[E_xE_y^2b-E^2b]+\frac{2}{N(N-1)}[E^2b-4E_xE_y^2b+2Eb^2].$$

Таким образом, J_N сходится к J в среднеквадратичном, и

$$\lim_{N\to\infty} NDJ_N=4\left[\int g^3(x)\,dx-\left(\int g^2(x)\,dx\right)^2\right]. \quad (8.6.20)$$

Далее, в случае ограниченности первой производной от $g(x)$, можно получить верхнюю границу для смещения J_N. Действительно,

$$|EJ_N-J|=\left|\int g(x)\left\{\int K(u)[g(x+uh_N)-g(x)]\,du\right\}dx\right|\leqslant$$
$$\leqslant\int g(x)\left\{\int K(u)\left|\frac{g(x+uh_N)-g(x)}{uh_N}\right|\cdot|u|\cdot h_N\,du\right\}dx\leqslant$$
$$\leqslant\sup_x|g'(x)|\int|u|K(u)\,du\cdot h_N=Ch_N,\ C<\infty.$$

8.6.2. КВАЗИ- U-СТАТИСТИКИ НА ГИСТОГРАММЕ

Другой класс оценок возникает, если в качестве непараметрической оценки плотности воспользоваться гистограммой. Квази-U-статистика I типа для функционала (8.6.2) будет иметь следующий вид:

$$J_N = \frac{1}{N}\sum_{j=1}^{N} F\left[g_N(x_j)\right] = \sum_k F\left[\frac{1}{N\Delta_k}\sum_{i=1}^{N}\pi(x_i, \Delta_k)\times\right.$$

$$\left.\times\frac{1}{N}\sum_{j=1}^{N}\pi(x_j, \Delta_k)\right] = \sum F\left[\frac{1}{N\Delta_k}\sum_{i=1}^{N}\pi(x_i, \Delta_k)\right]\times$$

$$\times\frac{1}{N}\sum_{j=1}^{N}\pi(x_j, \Delta_k) = \sum F\left[\frac{\tilde{p}_k}{\Delta_k}\right]\tilde{p}_k.$$

В отличие от оценок, рассмотренных в предыдущем разделе, в данном случае квази-U-статистики совпадают с прямыми оценками (сравни (8.6.22) с (8.5.27)) со всеми вытекающими отсюда последствиями (см. раздел 8.5.2).

Рассмотрим теперь оценки II типа. В этом случае квази-U-статистика имеет вид:

$$J_{N_1N_2} = \sum_k F\left[\frac{1}{N_1\Delta_k}\sum_{i=1}^{N_1}\pi(x_i, \Delta_k)\cdot\frac{1}{N_2}\sum_{j=1}^{N_2}(x_j, \Delta_k)\right] =$$

$$= \sum_k F\left[\frac{\tilde{p}_{1k}}{\Delta_k}\right]\cdot\tilde{p}_{2k}, \qquad (8.6.23)$$

где $\tilde{p}_{1k}$ и $\tilde{p}_{2k}$ относительные частоты (оценки p_k), полученные по подвыборкам объемов N_1 и N_2 соответственно. Учитывая асимптотическую нормальность и независимость $\tilde{p}_{1k}$ и $\tilde{p}_{2k}$ и предположив разложимость F в ряд Тэйлора, ограничившись (на тех же основаниях, что и в 8.5.2) двумя членами этого ряда, имеем:

$$FJ_{N_1N_2} \doteq \sum_k\left\{p_k F\left(\frac{p_k}{\Delta_k}\right) + F''\left(\frac{p_k}{\Delta_k}\right)p_k\cdot\frac{\sigma^2_{kN_1}}{\Delta_k^2}\right\},$$

откуда

$$\lim_{N_1\to\infty} EJ_{N_1N_2} = \sum_k p_k F\left(\frac{p_k}{\Delta_k}\right). \qquad (8.6.24)$$

Как видим, N_2 вообще не влияет на смещение. Это объясняется тем, что по второй подвыборке строится U-статистика, вообще не обладающая смещением (см. § 8.3). Проводя соответствующие вычисления для дисперсии $J_{N_1N_2}$, получим:

$$DJ_{N_1N_2} = \frac{1}{N_1}\left[\sum_k F_k^2 p_k - \left(\sum_k F_k p_k\right)^2\right] +$$

$$+\frac{1}{N_2}\left[\sum_k \gamma_k^2 p_k - \left(\sum_k \gamma_k p_k\right)^2\right], \qquad (8.6.25)$$

где $F_k = F\left(\frac{p_k}{\Delta_k}\right)$, $\gamma_k = F'\left(\frac{p_k}{\Delta_k}\right)\cdot\frac{p_k}{\Delta_k}$. Компоненты дисперсии

$J_{N_1N_2}$ за счет разных подвыборок имеют различный характер, что интересно само по себе. При $N_1=N_2=N/2$ (8.6.25) приведет к результату

$$\lim_{N\to\infty} NDJ_{\frac{N}{2}\frac{N}{2}}=2\left[\sum_k (F_k^2+\gamma_k^2)\,p_k-\left(\sum_k F_k p_k\right)^2-\left(\sum_k \gamma_k F_k\right)^2\right].$$

8.6.3. КВАЗИ- U-СТАТИСТИКИ НА ПОЛИГРАММЕ

Проделав все операции, необходимые для получения квази-U-статистики I типа, легко убедиться, что получатся статистики, отличающиеся от прямых оценок (8.5.32) на величину порядка N^{-1}. Это означает, что квази-U-статистики и прямые оценки на полиграмме асимптотически эквивалентны.

§ 8.7. ОЦЕНИВАНИЕ «НЕЯВНЫХ» ПАРАМЕТРОВ

До сих пор речь шла об оценивании параметров, приравниваемых к значению некоторого функционала. При этом оценка функционала и являлась оценкой соответствующего параметра. Однако в ряде случаев интересующий нас параметр θ_0 не может быть приравнен определенному функционалу. (Типичным примером является параметр запаздывания между двумя реализациями сигнала при разнесенном приеме). Довольно часто в таких случаях оцениваемый параметр удается связать с характером зависимости некоторого функционала $J(\theta_0, \theta)$ от произвольного параметра θ (той же природы и размерности, что и θ_0). Этот функционал задается таким образом, чтобы при $\theta=\theta_0$ и $J(\theta_0, \theta)$ обладал определенным, известным свойством: например, имел максимум (или минимум) или был равен некоторому конкретному значению (при монотонности по θ); и т. п. Тогда идея оценивания θ_0 состоит в том, чтобы принять в качестве $\hat{\theta}_0$ то значение θ, при котором оценка $J_N(\theta_0, \theta)$ обладает указанным свойством. Понятно, что качество получаемой таким образом оценки зависит от характера функционала J; от типа и качества оценки J_N; от свойств алгоритма поиска по θ; от типа или класса распределения выборки.

Имея в виду, что параметры такого типа задаются неявно, будем называть их неявными параметрами. Поскольку даже для одной задачи можно предложить несколько функционалов, обладающих требуемым свойством, возможно несколько вариантов оценок неявного параметра. Любопытно, что оценивание по методу максимума правдоподобия также

может быть интерпретировано как вариант изложенного метода. Пусть известна плотность $f(x/\theta)$. Определим энтропийный функционал

$$J(\theta_0, \theta) = -\int \ln f(x/\theta)\, dG(x/\theta_0), \qquad (8.7.1)$$

относительно которого известно (см., например, Тарасенко [4]), что он имеет единственный минимум при $f = dG/dx$. Функционал (8.7.1) допускает оценивание U-статистикой:

$$J_N = -\int \ln f(x/\theta)\, dG_N(x) = -\frac{1}{N}\sum_{i=1}^{N} \ln f(x_i/\theta), \qquad (8.7.2)$$

и нам остается лишь найти то значение θ, которое доставляет J_N минимальное значение. Таким образом, оценки максимального правдоподобия есть оценки параметров, неявно заданных через энтропийный функционал при известном типе плотности и использовании U-статистик. В чисто непараметрической постановке число возможных оценок неявного параметра увеличивается еще и за счет наличия разных способов оценивания одного и того же функционала (см. предыдущие параграфы данной главы).

Общий подход к оцениванию неявных параметров развит недостаточно и сводится пока к идее Ходжеса и Лемана [2], которые предложили (на примере оценки параметра сдвига) пользоваться непараметрическими тестами для проверки нулевой гипотезы против альтернативы, связанной с оцениваемым параметром (в их случае — со сдвигом). Проще всего продемонстрировать этот подход на примере двувыборочной задачи оценки сдвига.

Пусть имеются две выборки $x_1, \ldots, x_m$ и $y_1, \ldots, y_n$ из распределений $P(X \leqslant u) = F(u)$ и $P(Y \leqslant u) = F(u - \theta_0)$. Требуется оценить θ_0. Построим новую («сдвинутую») выборку $z_1, \ldots, z_n$, в которой $z_i = y_i + \theta$, $i = 1, \ldots, n$. Ясно, что при $\theta = \theta_0$ функции распределения величин X и Z совпадают. Поэтому при изменении θ некоторый тест однородности (см. § 9.13), отвергая гипотезу при больших $\theta - \theta_0$, в процессе сближения θ с θ_0, наконец, примет ее при некотором $\theta = \theta^* > \theta_0$. Аналогично, при сближении θ с θ_0 с другой стороны, мы получим некоторое значение $\theta^{**} < \theta$. Теперь в качестве оценки θ_0 мы можем взять величину $\hat{\theta} = (\theta^* + \theta^{**})/2$. Свойства этой оценки, естественно, будут связаны с качествами используемого теста.

В своей исходной статье Ходжес и Леман [2] рассмотрели свойства такой оценки сдвига при пользовании ранговыми тестами (см. гл. X), основанными на статистике Вилкоксона и статистике с нормальными метками. Интересно, что

при пользовании тестом Вилкоксона статистика θ оказалась эквивалентной медиане разностей $\{x_i - y_j\}$, а в одновыборочной задаче — медиане полусумм $(x_i + x_j)/2$, $i < j$. Этот факт исключает необходимость поисковой процедуры по θ в данном случае. Отметим, что аналогичная ситуация возникает при пользовании тестом Крамера — фон Мизеса (см. гл. IX). Мы снова приходим к медиане разностей (Файн [1]). Исследованию свойств оценок неявно задаваемого параметра сдвига посвящено большое количество работ. Появилось несколько статей по оценке параметра масштаба таким же методом, поскольку, деля выборочные значения одной из выборок на изменяемую величину θ, мы можем снова отыскать те значения θ^* и θ^{**}, при которых масштабы двух выборок станут неразличимыми для данного теста.

Однако законченная теория оценивания неявных параметров еще ждет своего построения. К числу вопросов, на которые должна ответить эта теория, относятся следующие: 1. Существует ли среди различных функционалов, через которые можно определить неявный параметр, наилучший, «наиболее чувствительный», т. е. дающий оценки наивысшего качества по заданному критерию? Если да, то как строить такой функционал для данной задачи? 2. Поскольку один и тот же функционал можно оценивать по-разному, то какой способ оценивания предпочтителен в смысле качества получаемой оценки параметра? 3. Случайным или закономерным является то, что в некоторых случаях оценка по методу Ходжеса-Лемана оказывается эквивалентной некоторой беспоисковой оценке? Если это закономерность, то нет ли способа нахождения сразу беспоисковой оценки? Выполнение такой программы является трудным, но чрезвычайно интересным делом.

ГЛАВА IX

КРИТЕРИИ СОГЛАСИЯ

§ 9.1. ЗАДАЧА СОГЛАСИЯ И ЕЕ ВАРИАНТЫ

Коротко задача согласия может быть сформулирована следующим образом: по независимым наблюдениям $x_1, \ldots, x_N$ требуется вынести суждение о том, принадлежат ли они к распределению заданного типа. Подобные задачи возникают в самых разнообразных случаях; типичным примером является необходимость экспериментальной проверки приемле-

мости принятой статистической модели. Хотя проблема проверки соответствия между теоретической моделью и экспериментальными данными является столь же древней, как и сама наука, по отношению к статистическим гипотезам эта проблема впервые была количественно поставлена и решена К Пирсоном [1] в 1900 г. В этой знаменитой работе, посвященной χ^2-критерию, Пирсон, в частности, продемонстрировал, сколь опасно переносить опыт сглаживания экспериментальных данных, содержащих лишь незначительные (по сравнению с основным эффектом) погрешности эксперимента, на аппроксимацию распределений случайных величин по выборке. Он не упустил возможности пошутить над двумя известными своими современниками — сэром Дж. Эйри и профессором Мерриманом. Оба они опубликовали по серии наблюдений, которые, как они утверждали, являются хорошим примером случайных величин, подчиняющихся нормальному распределению; суждение выносилось «на глаз», просто визуальным сравнением теоретической и экспериментальной кривых. Пирсон показал, что вероятность неслучайного соответствия этих кривых в случае Эйри равна 0,014; а у Мерримана оказалось всего полтора шанса на миллион! С этого момента и до настоящего времени задача согласия привлекает неослабевающее внимание статистиков. Различные варианты задачи согласия возникают в связи с тем, что в ее постановке могут фигурировать разные типы гипотез и альтернатив, а также различные предположения о наличном статистическом материале (одновыборочные, двувыборочные, многовыборочные задачи).

По типу гипотезы задачи согласия разбиваются на два класса.

Задача согласия с простой гипотезой. По выборке $x_1, \ldots, x_N$ независимых наблюдений из неизвестного распределения $G(x)$ проверить гипотезу $H_0: G(x) = F(x)$, где $F(x)$ заданная функция распределения. Непараметричность задачи выражается в непараметричности альтернативы, которая может быть сформулирована как в односторонних, так и в двустороннем вариантах:

$$\left.\begin{array}{l} H_1^-: F > G \\ H_1^+: F < G \end{array}\right\} \text{односторонние альтернативы} \qquad (9.1.1)$$

$$H_1: F \neq G \text{ двусторонняя альтернатива} \qquad (9.1.2)$$

Задача согласия со сложной гипотезой. Во многих случаях в статистической практике предлагаемая статистическая модель формулируется не в виде конкретного распределения, а в виде предположения о его типе: например,

предполагается, что выборочные значения распределены нормально с произвольными параметрами a и σ. Более общо, проверке подлежит гипотеза, состоящая в том, что распределение $F(x)$ принадлежит заданному параметрическому семейству функций $F(x; \theta_1, ..., \theta_k)$.

При формулировании типа альтернативы возможны также различные варианты задачи согласия. Так, если гипотеза состоит в том, что $F(x_1)=...=F(x_N)$, а альтернатива в том, что не все $F(x_i)$, $i=1, ..., N$, равны между собой, то задача называется задачей случайности. Если альтернатива состоит в том, что выборочные значения не являются независимыми, мы имеем так называемую задачу зависимости (которую мы будем рассматривать отдельно). Возможны также конкретные альтернативы регрессионного типа и т. д.

Далее отметим следующее существенное обстоятельство. Известно, что преобразование $u=F(x)$ в случае истинности распределения $F(x)$ приводит к тому, что случайная величина U оказывается равномерно распределенной в $[0, 1]$. В связи с этим возникает возможность приведения задачи согласия с простой гипотезой к так называемому каноническому виду (Чэпмен [1]):

$$H_0: F^*(u)=G^*(u)=FF^{-1}(u)=u \tag{9.1.3}$$

$$\left.\begin{array}{l} H_1^-: F^*(u)>G^*(u)=GF^{-1}(u) \\ H_1^+: F^*(u)<G^*(u)=GF^{-1}(u) \end{array}\right\} \text{ односторонние альтернативы} \tag{9.1.4}$$

$$H_1: F^*(u)\neq G^*(u) \text{ — двусторонняя альтернатива} \tag{9.1.5}$$

Здесь звездочками отмечены соответствующие распределения выборки в приведенном пространстве (см. гл. 1). Попутно заметим, что плотность вероятности приведенной выборки при гипотезе равна единице для всех $u\in[0, 1]$, альтернативная плотность приведенной выборки выражается через исходные плотности как

$$g^*(u)=\frac{g(F^{-1}(u))}{f(F^{-1}(u))}, \quad u\in[0, 1]. \tag{9.1.6}$$

Процедуры, предназначенные для решения задачи согласия, получили специальное название критериев согласия. Как и любые статистические тесты, критерии согласия имеют стандартную структуру: значение некоторой статистики $S(x_1, ..., x_N)$ сравнивается с определенным порогом (критическим значением) и в зависимости от того, превзойден или нет этот порог, принимается то или иное решение.

Специфика критерия согласия состоит прежде всего в том, что статистика S должна быть чувствительной к любым отличиям между истинным распределением $G(x)$ и гипотетическим распределением $F(x)$. В связи с этим первой проблемой, которая должна быть решена, является построение статистик, обладающих указанным свойством. Как мы увидим впоследствии, может быть предложено великое множество таких статистик; практически же полезными будут лишь те из них, при пользовании которыми можно установить однозначное соответствие между критическими значениями и уровнями значимости теста. Отсюда возникает вторая задача теории критериев согласия: найти точное или асимптотическое распределение данной статистики S при истинности гипотезы. Далее, было бы весьма желательным сравнить между собой многочисленные критерии согласия по их способности различать альтернативы того или иного типа. Поэтому третьей задачей теории критериев согласия является изучение их мощностных свойств при различных альтернативах. Все эти задачи будут обсуждаться в последующих параграфах данной главы.

§ 9.2. ПРИНЦИПЫ ПОСТРОЕНИЯ СТАТИСТИК ДЛЯ КРИТЕРИЕВ СОГЛАСИЯ

Поскольку задача согласия есть, по существу, задача сравнения двух распределений, одно или (в двувыборочном варианте) оба из которых могут быть неизвестными, то конструирование статистик для этих критериев опирается на три исходных пункта: а) выбирается та или иная мера различия между сравниваемыми распределениями («расстояние» между F и G); б) выбирается та или иная оценка неизвестного распределения по заданной выборке; в) тестовая статистика получается путем подстановки в функционал, выражающий «расстояние», выбранной оценки истинной функции распределения.

Многообразие критериев согласия объясняется прежде всего тем, что как оценки распределения, так и расстояния между ними могут быть введены различным образом. Кроме того, дополнительные классы критериев согласия появляются в связи с тем, что можно оценивать как функцию распределения, так и плотность вероятностей. Для построения критериев согласия могут быть использованы любые оценки функций распределения и плотностей; наиболее употребительные из них описаны в главе VII.

Что касается расстояний между двумя функциями распределения, то здесь также имеется широкий простор для

изобретательности. Известны, например, следующие варианты мер различия между функциями распределения (сводка которых имеется у Залера [1]):

$$d_1(F, G)=\sup_x [F(x)-G(x)]\,\psi(F(x)),$$

$$d_2(F, G)=\sup_x |F(x)-G(x)|\,\psi F(x)),$$

$$d_3(F, G)=\int_{-\infty}^{\infty} [F(x)-G(x)]\,dF(x),$$

$$d_4(F, G)=\int_{\infty}^{\infty} [F(x)-G(x)]^2\,\psi(F(x))\,dF(x),$$

$$d_5(F, G)=\int_{-\infty}^{\infty} \frac{F(x)}{G(x)}\,dF(x);$$

здесь через $\psi(t)\geqslant 0$ обозначена некоторая весовая функция. Аналогичным образом могут быть построены меры различия между плотностями. Например, ряд критериев согласия, использующих плотности в приведенном пространстве, построен (Кейл [1]) на «расстояниях» следующих типов:

$$d_6(F, G)=\int_0^1 \left\{\frac{d}{du}[GF^{-1}(u)-u]\right\}^2 du,$$

$$d_7(F, G)=\int_0^1 \left|\frac{d}{du}[GF^{-1}(u)-u]\right| du,$$

$$d_8(F, G)=-\int_0^1 \log\left[\frac{d}{du}GF^{-1}(u)\right] du.$$

Конечно, этим не исчерпывается многообразие возможных «расстояний».

Итак, построение тестовой статистики для любого критерия согласия может быть осуществлено путем подстановки в формулу для выбранного «расстояния» некоторой оценки соответствующего распределения. На деле, конечно, бывали случаи, когда тот или иной автор предлагал конкретную статистику из других, часто интуитивных, соображений; но впоследствии обычно оказывалось, что эту же статистику можно построить и указанным выше алгоритмическим способом.

§ 9.3. КРИТЕРИИ СОГЛАСИЯ НА СТАТИСТИКАХ «СТРУКТУРЫ d»

Общим для критериев этого класса является использование эмпирической функции распределения (см. § 7.2) $G_N(x)$

в качестве оценки $G(x)$. Получаемые при этом статистики оказываются симметричными относительно величин $F(x_1)$, ..., $F(x_N)$; для таких статистик введено специальное наименование статистик структуры d*. Различие критериев внутри этого класса достигается выбором конкретного «расстояния» $d(F, G)$. Многие из критериев этого класса достаточно хорошо изучены и поэтому мы ограничимся лишь упоминанием их основных свойств и ссылками на соответствующие работы.

1. Критерии Колмогорова-Смирнова предназначены для проверки гипотезы H_0 против альтернативы H_1 (D_N — критерий Колмогорова) и при альтернативах H_1^+ или H_1^- (D_N^+, D_N^- — критерии Смирнова) **. К статистикам Смирнова приводит использование расстояния $d_1(F, G)$ с $\psi(t)=1$. Эти статистики записываются в виде:

$$D_N^+ = \sup_x [G_N(x)-F(x)] = \max_{1\leqslant i\leqslant N}\left[\frac{i}{N}-F(x_{(i)})\right],$$

$$D_N^- = -\inf_x [G_N(x)-F(x)] = \max_{1\leqslant i\leqslant N}\left[F(x_{(i)})-\frac{i-1}{N}\right]. \quad (9.3.1)$$

Для статистики Колмогорова выбрано расстояние $d_2(F, G)$ с $\psi(t)\equiv 1$, и выглядит она как

$$D_N = \sup_x |G_N(x)-F(x)| = \max [D_N^+, D_N^-]. \quad (9.3.2)$$

Во многих работах приводятся распределения этих статистик; за подробностями мы отсылаем читателя к обзорным статьям Дарлинга [1] и Залера [1]. Их асимптотические распределения не являются нормальными. Например, обе статистики D_N^+ и D_N^- асимптотически имеют распределение

$$\lim_{N\to\infty} P\{D_N^+\sqrt{N}\leqslant d\} = 1-e^{-2d^2},$$

$$0\leqslant d\leqslant\infty.$$

Распределения этих статистик табулированы Бирнбаумом [1], Оуэном [1], Большевым и Смирновым [1].

---

* Характерным свойством статистик структуры d является то, что при $F=G$ (т. е. при истинности гипотезы) их распределение не зависит от F. Это есть простое следствие того, что величины $F(x)$ при $F=G$ всегда распределены равномерно. Хотя этот факт почти тривиален, многие авторы, желая подчеркнуть его значение, формулируют его в виде теоремы о непараметричности (distribution-free) статистик структуры d.

** Здесь и в дальнейшем для удобства критерий, основанный на статистике S, будем называть S-критерием.

2. Критерии Реньи. Статистики этих критериев соответствуют выбору «расстояний» $d_1(F, G)$ и $d_2(F, G)$ с определенными весовыми функциями $\psi(t)$. Выбор весовой функции $\psi(t)$ в виде

$$\psi[F(x)]=\begin{cases}\dfrac{1}{F(x)}, & F(x)\geqslant b,\\ 0, & F(x)<b,\end{cases}$$

где b — заранее данное число, $0\leqslant b\leqslant 1$, приводит к статистикам, задаваемым следующими формулами:

$$R_N^+(b, 1)=\sup_{F(x)\geqslant b}\frac{G_N(x)-F(x)}{F(x)}=\max_{F(x_{(i)})\geqslant b}\frac{\frac{i}{N}-F(x_{(i)})}{F(x_{(i)})},$$

$$R_N^-(b, 1)=\inf_{F(x)>b}\frac{G_N(x)-F(x)}{F(x)}=\max_{F(x_{(i)})\geqslant b}\frac{F(x_{(i)})-\frac{i-1}{N}}{F(x_{(i)})}, \quad (9.3.3)$$

$$R_N(b, 1)=\sup_{F(x)\geqslant b}\frac{|G_N(x)-F(x)|}{F(x)}=\max[R_N^+(b, 1), R_N^-(b, 1)].$$

Нетрудно видеть, что эти статистики являются также статистиками «структуры d».

Если же весовая функция определяется равенством

$$\psi[F(x)]=\begin{cases}\dfrac{1}{1-F(x)}, & F(x)\leqslant b,\\ 0, & F(x)>b,\end{cases}$$

то статистики критериев задаются выражениями

$$R_N^+(0, b)=\sup_{F(x)<b}\frac{G_N(x)-F(x)}{1-F(x)}=\max_{F(x_{(i)})<b}\frac{F(x_{(i)})-\frac{i-1}{N}}{1-F(x_{(i)})},$$

$$R_N^-(0, b)=-\inf_{F(x)\leqslant b}\frac{G_N(x)-F(x)}{1-F(x)}=\max_{F(x_{(i)})<b}\frac{F(x_{(i)})-\frac{i-1}{N}}{1-F(x_{(i)})}, \quad (9.3.4)$$

$$R_N(0, b)=\sup_{F(x)<b}\frac{|G(x)-F(x)|}{1-F(x)}=\max[R_N^+(0, b), R_N^-(0, b)].$$

Статистики $R_N^+(b, 1)$, $R_N^-(b, 1)$, $R_N^+(0, 1-b)$, $R_N^-(0, 1-b)$ имеют одинаковые распределения (аналогичное утверждение справедливо и для статистик $R_N(b, 1)$ и $R_N(0, 1-b)$, поэто-

му для вычисления критических значений соответствующих статистик достаточно знать распределения статистик $R_N^+(b, 1)$ и $R_N(b, 1)$.

Асимптотические распределения этих статистик были найдены Реньи [1]. При $0<b\leqslant 1$ имеет место предельное соотношение

$$\lim_{N\to\infty} P\left\{\sqrt{\frac{Nb}{1-b}}\, R_N^+(b, 1)<x\right\}=2\Phi(x)-1, \quad x<0,$$

где $\Phi(x)$ — стандартное нормальное распределение. Асимптотическое распределение статистики $R_N(b, 1)$ записывается в виде

$$\lim_{N\to\infty} P\left\{\sqrt{\frac{Nb}{1-b}}\, R_N(b, 1)<x\right\}=L(x), \quad x>0,$$

где

$$L(x)=\frac{4}{\pi}\sum_{k=0}^{\infty}\frac{(-1)^k}{2k+1}\cdot e^{-\frac{(2k+1)^2\pi^2}{8x^2}}.$$

Таблицы этих распределений можно найти у Большева и Смирнова [1] и у Оуэна [1].

3. Критерии на интегральных «расстояниях». Эти критерии, в отличие от рассмотренных выше, соответствуют выбору «расстояний» в интегральной форме. Так, например, выбрав расстояние $d_3(F, G)$, получим статистику, предложенную Мозесом:

$$T_N^-=N\int_{-\infty}^{\infty}[G_N(x)-F(x)]\,dF(x)=\frac{1}{N}\sum_{i=1}^{N}F(x_i). \quad (9.3.5)$$

Статистика Мозеса также является статистикой структуры d. Статистика T_N^- при гипотезе H_0 имеет асимптотически нормальное распределение со средним значением, равным $^1/_2$ и дисперсией $^1/_{12}N$.

Другой статистике, предложенной Пирсоном [1], соответствует выбор «расстояния» $d_5(F, G)$:

$$\pi_N^-=2N\int_{-\infty}^{\infty}\frac{G_N(x)}{F(x)}\,dF(x)=-2\sum_{i=1}^{N}\ln F(x_i). \quad (9.3.6)$$

При гипотезе H_0 эта статистика имеет χ^2-распределение с $2N$ степенями свободы; π_N^--критерий называют иногда критерием произведения вероятностей Пирсона. Критическая область π_N^--теста охватывает малые значения статистики.

Критерии Андерсона и Дарлинга соответствуют выбору «расстояния» $d_4(F, G)$. Статистики этих критериев, предна-

значенных для проверки гипотезы H_0 против H_1, записываются в таком виде:

$$\int_{-\infty}^{\infty} [G_N(x)-F(x)]^2 \psi[F(x)]\, dF(x)=$$

$$=\frac{2}{N}\sum_{i=1}^{N}\left\{\varphi_1[F(x_{(i)})]-\frac{2i-1}{2N}\varphi_2[F(x_{(i)})]\right\}+\int_0^1 (1-t)^2\psi(t)\,dt,$$

где

$$\varphi_1(t)=\int_0^t x\psi(x)\,dx,\quad \varphi_2(t)=\int_0^t \psi(x)\,dx.$$

В частности, при $\psi(t)\equiv 1$ получим статистику Крамера — фон Мизеса — Смирнова

$$N\omega_N^2=\frac{1}{12N}+\sum_{i=1}^{N}\left[F(x_{(i)})-\frac{2i-1}{2N}\right]^2; \qquad (9.3.7)$$

при $\psi(t)=1/t(1-t)$ получим статистику Андерсона и Дарлинга [1]

$$\Omega_N^2=-N-2\sum_{i=1}^{N}\left\{\frac{2i-1}{2N}\ln F(x_{(i)})+\right.$$

$$\left.+\left(1-\frac{2i-1}{2N}\right)\ln[1-F(x_{(i)})]\right\}. \qquad (9.3.8)$$

Эти статистики также являются статистиками структуры d и, следовательно, их распределения не зависят от $F(x)$ при гипотезе. Асимптотическое распределение статистики $N\omega_N^2$, которое оказывается не нормальным при гипотезе H_0, впервые было получено Смирновым [3]; позднее Андерсон и Дарлинг [1] получили более точный результат и табулировали величину

$$\lim_{N\to\infty} P(N\omega_N^2<\omega).$$

Распределение величины ω_N^2 при малом объеме выборки было исследовано Пирсоном и Стефенсом [1], которые показали, что сходимость к предельному распределению с увеличением N является очень быстрой. Для выборок объема $N>20$, асимптотическое распределение отвечает всем практическим требованиям к точности.

Модификация статистики $N\omega_N^2$, предложенная Андерсоном и Дарлингом, сводится к введению весовой функции $\psi(t)=1/t(1-t)$. Такая модификация приводит к статистикам, более чувствительным к отклонениям на «хвостах» распределений. Этими авторами было получено асимптотическое

распределение статистики Ω_N^2, а Льюис [1] табулировал его и исследовал распределение для малых объемов выборки. Оказывается, что сходимость к предельному распределению, как и для статистики $N\omega_N^2$, является чрезвычайно быстрой, и уже при $N>10$ асимптотические результаты дают удовлетворительную аппроксимацию.

Укажем на один метод, позволяющий в ряде случаев быстро получить асимптотические распределения тестовых статистик структуры при гипотезе. Этот метод основан на представлении тестовых статистик в виде некоторого случайного процесса, как это впервые проделал Дуб [1]. Введем процессы

$$\eta_N(x)=\sqrt{N}\,[F_N(x)-F(x)]$$

и

$$\xi_N(t)=\eta_N(F^{-1}(t))=\sqrt{N}\,[F_N F^{-1}(t)-t]. \tag{9.3.9}$$

В силу леммы 1 (§ 8 гл. IX, Гихман и Скороход [1]) последовательность случайных процессов $\xi_N(t)$ при $N\to\infty$ слабо сходится к гауссову процессу $\xi(t)$, для которого $M\xi(t)=0$, $M[\xi(t)\cdot\xi(s)]=t(1-s)$, $0\leqslant t<s\leqslant 1$. Очевидно, тестовые статистики данного параграфа могут быть выражены через случайный процесс $\xi_N(t)$. Например,

$$\sqrt{N}D_N^+=\sup_{0<t<1}\xi_N(t); \tag{9.3.10}$$

$$\sqrt{N}D_N=\sup_t|\xi_N(t)|; \tag{9.3.11}$$

$$\sqrt{\frac{Nb}{1-b}}\,R_N^+(b,1)=\sup_{t\in[0,1]}[\xi_N(t)\psi(t)], \tag{9.3.12}$$

где

$$\psi(t)=\sqrt{\frac{b}{1-b}}\begin{cases}\dfrac{1}{t}, & t\geqslant b,\\ 0, & t<b;\end{cases}$$

$$\omega_N^2=\int_0^1\xi_N^2(t)\,dt; \tag{9.3.13}$$

$$T_N^-=\int_0^1\xi_N(t)\,dt. \tag{9.3.14}$$

Согласно лемме 2 (§ 8 гл. IX, Гихман и Скороход [1]),

$$\lim_{N\to\infty}P\{\varphi(\xi_N(t))<x\}=P\{\varphi(\xi(t))<x\},$$

где φ — непрерывный на $D_{[0,1]}$ функционал (приведенные выше представления статистик удовлетворяют этому свойству). Последнее соотношение и позволяет находить асимптотические распределения соответствующих тестовых статистик. Примеры, иллюстрирующие такой подход, смотри у Гихмана и Скорохода [1], Конюховского [1], Гельцера и Пайка [1], Дуба [1] и др.

§ 9.4. КРИТЕРИИ СОГЛАСИЯ НА ВЫБОРОЧНЫХ ИНТЕРВАЛАХ ПРИВЕДЕННОЙ ВЫБОРКИ

Если при построении тестовой статистики пользоваться оценками распределений, основанными на выборочных интервалах (т. е. полиграммами или их интегралами), то возникает класс критериев согласия, свойства которых связаны со статистическими свойствами выборочных интервалов. Дадим краткий обзор данного класса критериев согласия, заметив при этом, что многие тестовые статистики этого типа были предложены чисто эвристически, так что пока не всегда можно указать, какому «расстоянию» соответствует данная статистика.

Будем рассматривать далее лишь критерии, основанные на выборочных интервалах приведенной выборки. Пусть $x_1, \ldots, x_N$ — выборочные значения, приведенные в интервал $[0, 1]$ с помощью некоторой функции распределения F (см. § 1.6). Выборочные интервалы определяются как $\Delta_R = x_{(R+1)} - x_{(R)}$, $R = 1, 2, \ldots, N$, $x_{(0)} = 0$, $x_{(N+1)} = 1$. Статистики $S(\Delta_0, \ldots, \Delta_N)$, симметричные относительно своих переменных, будем называть статистиками «структуры D» (Тарасенко [7]). Поскольку при совпадении функции преобразования F с истинной функцией распределения G выборочные интервалы $\{\Delta_R\}$ распределены известным образом независимо от типа распределения (см. § 4.3), статистики структуры D, на том же основании, что и статистики структуры D, могут быть названы непараметрическими (distribution — free). Статистические свойства статистик структуры D при гипотезе могут быть получены с помощью теории случайного разбиения единичного отрезка (Моран [1, 2], Кимбалл [1] и др.).

К настоящему времени наиболее детальные результаты получены для тех критериев согласия, которые основаны на полиграмме первого порядка (см. § 7.5); напомним, что соответствующая оценка функции распределения выразится в этом случае как *

---

* Впервые оценка функции распределения (9.4.1) была предложена Пайком [2].

$$G_N(x)=\frac{R}{N+1}+\frac{x-x_{(R)}}{(N+1)\Delta_R},\quad x\in\Delta_R,\quad R=1,2,\ldots,N+1. \tag{9.4.1}$$

Кроме того, эвристически были предложены статистики, использующие функции выборочных интервалов, например, их отношения, упорядоченные интервалы, их частичные суммы и т. д.

Наибольшее внимание исследователей привлекли статистики, имеющие аддитивную структуру, т. е. имеющие вид

$$T=\sum_{R=1}^{N+1}h(\Delta_R), \tag{9.4.2}$$

где $h(x)$ — некоторая функция, определенная на $[0,1]$.

Общий метод нахождения распределений статистик (9.4.2) при гипотезе был предложен Дарлингом [2]. Им было найдено выражение для характеристической функции статистики T при произвольной функции $h(x)$:

$$Ee^{i\xi T}=\frac{N!}{2\pi i}\int_{C-i\infty}^{C+i\infty}e^{y}\Big(\int_0^\infty e^{-xy+i\xi h(x)}\,dx\Big)^{N+1}dy. \tag{9.4.3}$$

Используя это выражение, нетрудно получить моменты статистики

$$\mu_1=ET=N(N+1)\int_0^1(1-x)^{N-1}h(x)\,dx,$$

$$\mu_2=ET^2=N(N+1)\int_0^1(1-x)^{N-1}h^2(x)\,dx+$$

$$+N^2(N^2-1)\iint\limits_{\substack{0<x+y<1\\ x>0,\ y>0}}(1-x-y)^{N-2}h(x)h(y)\,dx\,dy. \tag{9.4.4}$$

(Естественно, эти же формулы получатся при непосредственном использовании распределений § 4.3). Исследуя асимптотическое поведение характеристической функции (9.4.3) при заданной $h(x)$, можно установить асимптотическое распределение данной статистики.

Приведем теперь результаты для наиболее изученных статистик вида (9.4.2).

Статистики типа

$$T_1^*=\sum_{i=1}^{N+1}\Delta_i^\alpha,\quad \alpha\neq 1 \tag{9.4.5}$$

получили название статистик Кимбалла. Изучение этих статистик началось со случая $\alpha=2$ (Гринвуд [1]); Моран [1] доказал асимптотическую нормальность T_1^2 при ги-

потезе; в дискуссии по статье Гринвуда Ирвин предложил рассматривать статистику $\sum\left(\Delta_R-\frac{R}{N+1}\right)^2$, асимптотическая нормальность которой при гипотезе была доказана Кимбаллом [1]. (Заметим кстати, что статистики Ирвина и Гринвуда эквивалентны, поскольку $\Sigma\Delta_R=1$). Случай произвольного $\alpha>1$ наиболее полно рассматривался Дарлингом [2], который, исследуя асимптотическое поведение характеристической функции (9.4.3) при $h(x)=x^\alpha$, показал асимптотическую нормальность статистик Кимбалла T_1^α. ($T_1^\alpha\sim N(\mu_N, \sigma_N)$). Параметры μ_N и σ_N соответственно определяются выражениями

$$\mu_N\sim\frac{\Gamma(\alpha+1)}{N^{\alpha-1}},$$

$$\sigma_N\sim\frac{1}{N^{2\alpha-1}}[\Gamma(2\alpha+1)-(\alpha^2+1)\,\Gamma^2(\alpha+1)], \qquad (9.4.6)$$

где $\Gamma(x)$ — гамма-функция. Таким образом, при гипотезе ($f(x)=1$, $x\in[0, 1]$) распределение случайной величины

$$\frac{N^{\alpha-1/2}\cdot T_1^\alpha-\sqrt{N}\,\Gamma(\alpha+1)}{[\Gamma(2\alpha+1)-(\alpha^2+1)\,\Gamma^2(\alpha+1)]^{1/2}} \qquad (9.4.7)$$

стремится при $N\to\infty$ к стандартному нормальному распределению.

Следующий тип статистик (9.4.2) можно получить, используя «расстояние»

$$d=\int_0^1\left|\frac{d}{du}[G(u)-u]\right|du \qquad (9.4.8)$$

и оценку $G(u)$ в виде (9.4.1). В результате получится статистика Кендалла—Шермана

$$T_2=\frac{1}{2}\sum_{R=1}^{N+1}\left|\Delta_R-\frac{1}{N+1}\right|, \qquad (9.4.9)$$

которая была предложена Кендаллом (в дискуссии по статье Гринвуда [1]) и детально изучена Шерманом [1]. Последний доказал асимптотическую нормальность статистики T_2 при гипотезе и нашел выражение для ее моментов также при истинности гипотезы:

$$E(T_2)^\mu=\binom{N+\mu}{\mu}\sum_{s=0}^{\nu-1}\binom{\mu-1}{s}\left(\frac{N-s}{N+1}\right)^{N+\mu}. \qquad (9.4.10)$$

В частности, среднее и дисперсия T_2 соответственно равны

$$ET_2=\left(\frac{N}{N+1}\right)^{N+1}\longrightarrow e^{-1}, \tag{9.4.11}$$

$$DT_2=\frac{2N^{N+2}+N(N-1)^{N+2}}{(N+2)(N+1)^{N+2}}-\left(\frac{N}{N+1}\right)^{2N-2}\longrightarrow$$

$$\longrightarrow\frac{2e-5}{e^2}\cdot\frac{1}{N}. \tag{9.4.12}$$

Следовательно, распределение случайной величины

$$\left(\frac{Ne^2}{2e-5}\right)^{1/2}\left(T_2-\frac{1}{e}\right)$$

стремится при $N\to\infty$ к стандартному нормальному распределению при истинности гипотезы.

Рассматривая различные возможности определения функции $h(x)$ в (9.4.2), Дарлинг [2] предложил использовать функции $h(x)=\ln x$ и $h(x)=\frac{1}{x}$. К статистикам такого вида приводят соответственно «расстояния»

$$d=-\int_0^1 \ln\frac{du}{dG}\,dG, \tag{9.4.13}$$

$$d=\int_0^1\left\{\frac{d}{du}[G(u)-u]\right\}^2 du. \tag{9.4.14}$$

Запишем статистики Дарлинга в явном виде

$$T_3=\sum_{R=1}^{N+1}\ln\Delta_R, \tag{9.4.13'}$$

$$T_4=\sum_{R=1}^{N+1}\frac{1}{\Delta_R}. \tag{9.4.14'}$$

Как показал Дарлинг, T_3* при гипотезе имеет асимптотически нормальное распределение со средним $ET_3=-(N+1)(\ln N-C)$ и дисперсией $DT_3=(N+1)(\frac{\pi^2}{6}-1)$. (Здесь $C=0{,}577\ldots$ — константа Эйлера). Асимптотическое распределение T_4 не является нормальным в связи с неограниченностью моментов величин $1/\Delta_R$ (см. § 4.6).

При использовании в качестве «расстояния» кульбаковской дивергенции информации

---

* Статистика рассматривалась также в работах Кейла и Годамбе [1], Тарасенко [3].

$$d=-\int_0^1 \ln\left(\frac{dG}{du}\right)du \qquad (9.4.15)$$

получается статистика Кейла—Геберта

$$T_5=\sum_{R=1}^{N+1}\Delta_R\ln\Delta_R+\ln(N+1), \qquad (9.4.15')$$

предложенная Кейлом [2]. Геберт и Кейл [1] показали асимптотическую нормальность T_5 при истинности гипотезы; среднее и дисперсия T_5 равны соответственно $ET_5=1-C$ и

$$DT_5=\frac{1}{N+1}\left(\frac{\pi^2}{3}-1\right).$$

Особый класс тестовых статистик возникает при использовании упорядоченных выборочных интервалов. Основанием для этого служит тот факт, что при отклонении распределения от равномерного должно происходить (статистическое) уменьшение малых интервалов и увеличение больших. Естественно поэтому, что прежде всего уделялось внимание таким статистикам, как

$$T_6=\max_R\Delta_R=\Delta_{(N+1)} \text{ и } T_7=\min_R\Delta_R=\Delta_{(1)}. \qquad (9.4.16)$$

Статистики (9.4.16) можно получить и алгоритмически, если взять в качестве «расстояний» между распределениями выражения типа (Кейл [1])

$$d=\sup_x\left|\frac{dG}{dx}-1\right|,$$

$$d^+=\sup_x\left(\frac{dG}{dx}-1\right),$$

$$d^-=\sup_x\left(1-\frac{dG}{dx}\right).$$

В дискуссии по статье Гринвуда [1] Кендалл предложил также использовать статистики

$$T_8=\Delta_{(N+1)}-\Delta_{(1)} \text{ и } T_9=\Delta_{(N+1)}/\Delta_{(1)}. \qquad (9.4.17)$$

Однако позднее Вайсс [2] показал, что асимптотически статистика T_8 эквивалентна T_6, а T_9 эквивалентна T_7. Далее, Дарлинг [2] показал, что статистики T_6 и T_7 при истинной гипотезе асимптотически независимы и экспоненциально распределены:

$$\lim_{N\to\infty}P\left\{T_7>\frac{u}{(N+1)^2},\quad T_6<\frac{\ln(N+1)-\ln v}{N+1}\right\}=e^{-(u+v)},$$

$(u>0,\ v>0)$, что, впрочем, является проявлением общих свойств выборочных интервалов (см. § 4.5).

Так как статистическому увеличению (уменьшению) при отклонении от гипотезы подвержен не только максимальный (минимальный) интервал, естественным обобщением является введение частичных сумм упорядоченных интервалов, т. е. статистик вида

$$T_{10}=\Delta_{(1)}+\ldots+\Delta_{(k)},\quad T_{11}=\Delta_{(N-k)}+\ldots+\Delta_{(N+1)}.$$

Такие статистики изучались Мулдоном [1], Блюменталем [1], Бартоном и Дэвидом [1] и другими.

§ 9.5. χ^2-КРИТЕРИЙ

χ^2-критерий Пирсона [1] является исторически первым критерием согласия и по настоящее время одним из наиболее употребляющихся в практике. Характерной чертой этого критерия является то, что он направлен на обнаружение различий между гипотетическим и эмпирическим распределениями вероятностей (для дискретных величин) или соответствующими плотностями (для непрерывного случая).

Предположим сначала, что мы имеем дело с дискретной случайной величиной Y, принимающей k значений $Y_1, \ldots, Y_k$, относительно которых выдвигается предположение, что их вероятности равны соответственно $P(Y_1)=p_1, \ldots, P(Y_k)=p_k$. Требуется по выборке объема N проверить, согласуется ли результат эксперимента с предположением о величинах $\{p_i\}$, $i=1, \ldots, k$. По гипотезе определяются средние значения $\{m_i\}$ чисел осуществления величин $\{Y_i\}$: $m_1=Np_1, \ldots, m_k=Np_k$. Пусть в эксперименте величины Y_i осуществились n_i раз. Задав некоторую меру различия между $\{m_i\}$ и $\{n_i\}$, получим тот или иной критерий согласия между ними. Пирсон предложил для этого величину

$$X^2=\sum_{i=1}^{k}\frac{(n_i-m_i)^2}{m_i}=\sum_{i=1}^{k}\frac{n_i^2}{m_i}-N. \qquad (9.5.1)$$

Хорошо известно, что случайная величина X^2 асимптотически распределена при гипотезе согласно распределению χ^2:

$$f(\chi^2)\,d\chi^2=\frac{1}{2^{\frac{\nu}{2}}\left(\frac{\nu}{2}-1\right)!}(\chi^2)^{\frac{\nu}{2}-1}e^{-\frac{\chi^2}{2}}\,d\chi^2, \qquad (9.5.2)$$

где ν — число степеней свободы. Этому же распределению подчиняется сумма квадратов ν независимых нормальных случайных с нулевым средним и единичной дисперсией. Для ста-

тистики (9.5.1) число степеней свободы равно $k-1$. При малых объемах выборки (9.5.2) выполняется лишь приближенно и поправочные члены зависят от конкретной гипотезы $\{P_i\}$; поэтому χ^2-критерий является лишь асимптотически непараметрическим (Кокран [1]).

χ^2-критерий широко применяется и для проверки гипотез о характере непрерывных распределений. Техника его применения основывается на возможности представления непрерывной величины некоторым ее дискретным эквивалентом. Достигается это разбиением интервала возможных значений случайной величины X на непересекающиеся и прилегающие друг к другу интервалы группировки $\Delta x_i = x_{i+1} - x_i$, где $\{x_i\}$ — совокупность пороговых значений и последующего рассмотрения дискретной случайной величины Y со значениями $\{Y_i : Y_i \equiv X \in \Delta x_i\}$. Гипотетическое распределение введенной таким образом случайной величины выразится, очевидно, как

$$p_i = \int_{\Delta x_i} p(x)\,dx, \quad i = \ldots, -1, 0, 1, \ldots,$$

где $p(x)$ — гипотетическая плотность вероятностей непрерывной величины X. Очевидно, величины, входящие в (9.5.1), выразятся как

$$m_i = Np_i; \quad n_i = \sum_{j=1}^{N} [C(x_i - x_j) - C(x_{i+1} - x_j)].$$

После этого проводится стандартное применение χ^2-критерия с применением соответствующих таблиц (см., например, Большев и Смирнов [1]).

Как подчеркнул Кокран [1], использование χ^2-критерия для непрерывных плотностей требует учета ряда особенностей и тонкостей, которым нередко не уделяется должного внимания. Прежде всего это связано с выбором числа, длин и размещения интервалов группировки $\{\Delta x_i\}$. Приведем сразу перечень рекомендаций, основанный на детальном рассмотрении этого вопроса Манном и Вальдом [1], Гумбелем [2], Вильямсом [1] и другими. (Подробности смотри у Кокрана [1]).

1. В отличие от обычной практики, интервалы группировки $\{\Delta x_i\}$ для гистограммы должны выбираться не равными, а так, чтобы числа m_i были по возможности одинаковыми.

2. Мощность χ^2-критерия будет зависеть от числа k интервалов группировки и, конечно, от конкретной альтернативы. Имея в виду непараметрическую альтернативу, Манн и Вальд [1] характеризуют различие между гипотезой и альтернативой величиной $\delta = \sup_x |F(x) - G(x)|$. Для этого достаточно общего случая найдено значение k, наилучшее в

том смысле, что при любой конкретной альтернативе из этого класса мощность теста будет не меньше $^1/_2$. Это значение k определяется формулой

$$k=4\left[\frac{2(N-1)^2}{C^2}\right]^{1/5}, \qquad (9.5.3)$$

где C определяется из соотношения

$$\frac{1}{\sqrt{2\pi}}\int_C^\infty e^{-\frac{x^2}{2}}dx=\alpha$$

и α — уровень значимости критерия. Для ориентировки приведем таблицу значений k в зависимости от N при $\alpha=0{,}05$ (табл. 9.5.1).

Таблица 9.5.1

N	200	400	600	800	1000	1500	2000
k	31	41	48	54	59	70	78

Вильямс [1], однако, показал, что мощность χ^2-критерия не слишком критична к выбору k и на практике число интервалов можно без особых потерь уменьшить примерно вдвое.

3. Обычно считается, что для успешного применения χ^2-критерия достаточно, чтобы в каждом из интервалов группировки было не менее 5—6 выборочных значений. Однако из вышеприведенной таблицы следует, что для N от 200 до 1000 k меняется от 6 до 16, а с учетом поправки Вильямса — от 12 до 32, что значительно превышает рекомендуемую обычно цифру.

§ 9.6. О СРАВНЕНИИ КРИТЕРИЕВ СОГЛАСИЯ СТРУКТУРЫ d ПО ИХ МОЩНОСТИ

Наличие большого числа критериев согласия неизбежно ставит в каждом конкретном случае задачу выбора одного из них, то есть задачу сравнения критериев по их качеству. Характеристики качества, естественно, могут быть различными: иногда решающим является простота выполнения процедуры; иногда все зависит от того, имеются ли под рукой соответствующие таблицы; в ряде случаев предпочтение должно быть отдано тому критерию, который наиболее чувствителен к отличиям альтернативы от гипотезы.

При параметрической формулировке простой альтернативы тесты, построенные на отношении правдоподобия, являют-

ся, очевидно, равномерно наиболее мощными: структура таких критериев согласия может быть выражена соотношением

$$\sum_{i=1}^{N} \ln g^*(F(x_i)) > C,$$

где g^* — альтернативная плотность приведенной выборки, F — гипотетическая функция распределения выборки, C — граница критической области. Свойства таких тестов согласия подробно исследованы Беллом и Доксумом [1].

Если же альтернатива непараметрична (а это — наиболее типичная ситуация, в которых применяются критерии согласия), то задача сравнения тестов по их мощности теряет однозначность, которую в некоторой степени можно устранить, рассматривая определенные классы альтернатив, верхнюю или нижнюю границы мощностей внутри этих классов. Другую возможность предоставляет рассмотрение «контигуальных» альтернатив (то есть последовательности альтернатив, определенным образом сходящихся к гипотезе); другими словами, критерии согласия можно сравнивать по их питмановской или бахадуровской эффективностям (см. § 2.7–2.9). Обсудим вкратце некоторые результаты, полученные в этих направлениях; укажем сразу, что полученные к настоящему времени результаты пока трудно увязать в единую стройную систему из-за разнобоя исходных предпосылок различных авторов.

Достаточно общий подход предложен Бирнбаумом [2] и развит Чэпменом [1]. Этот подход состоит в том, чтобы задать некоторый непараметрический класс альтернатив таким образом, который позволил бы выделить в этом классе «наилучшую» и «наихудшую» альтернативы, при которых достигается верхняя и, соответственно, нижняя границы мощности данного теста. Такой класс альтернатив оказывается возможно ввести лишь для частично упорядоченных тестов (partially ordered tests), ориентированных на одностороннюю задачу согласия (Чэпмен [1]).

Определение 9.6.1. Критерий называется частично упорядоченным, если из $G_1(x) \leqslant G_2(x)$ следует $\beta(G_1) \geqslant \beta(G_2)$. Через $\beta(G)$ обозначена мощность критерия при альтернативе G.

Определение 9.6.2. Критерий называется монотонным, если из $x_i \geqslant x_i^*$ $(i=1, \ldots, N)$ следует $S(x_1, \ldots, x_N) \geqslant S(x_1^*, \ldots, x_N^*)$. Через S обозначена тестовая статистика.

Теорема Чэпмена. *Монотонный тест структуры d является частично упорядоченным и несмещенным.* (Доказательство достаточно просто и приведено у Чэпмена [1]).

Рассмотрим сначала случай, когда «расстояние» между альтернативой и гипотезой вводится в виде

$$\delta = \sup_u [G^*(u) - u] = \sup_x (G(x) - F(x)). \qquad (9.6.1)$$

Исходя из приведенных выше определений и теоремы, легко указать наиболее и наименее благоприятные альтернативы (рис. 9.6.1), дающие верхнюю (при G_B) и нижнюю (при G_H) границы мощности для одностороннего, частично упорядоченного теста при фиксированном расстоянии δ. Альтернативы

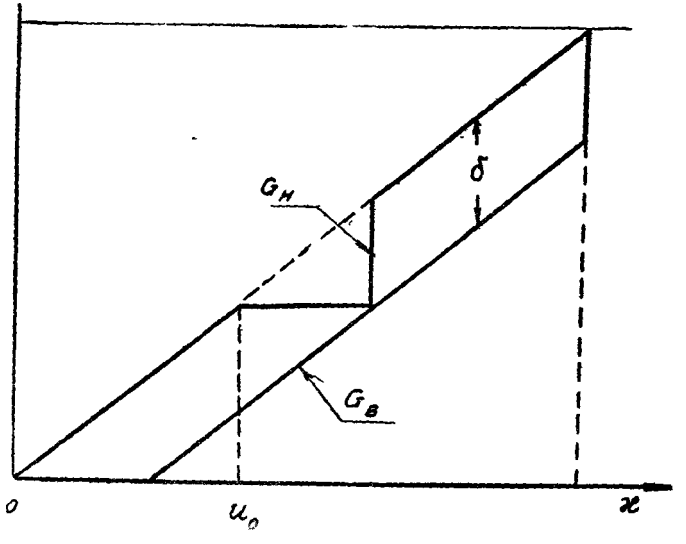

Рис. 9.6.1

заданы таким образом, чтобы мощность можно было вычислить, пользуясь результатами, известными для гипотезы, так как G_B и G_H являются комбинацией равномерных и дельта-образных распределений. Поскольку G_H имеет произвольный параметр u_0 (см. рис. 9.6.1), то нижнюю границу мощности необходимо минимизировать по этому параметру. С учетом этого замечания, выражения для верхней ($\overline{\beta}_S(\delta)$) и нижней ($\underline{\beta}_S(\delta)$) границ мощности запишутся следующим образом:

$$\overline{\beta}_S(\delta) = \beta_S(G_B), \qquad (9.6.2)$$

$$\underline{\beta}_S(\delta) = \min_{0 \leqslant u_0 \leqslant 1-\delta} \beta_S(G_H), \qquad (9.6.3)$$

здесь через $\beta_S(G_H)$ обозначена мощность S-критерия при альтернативе G.

В качестве простого примера рассмотрим нахождение границ мощности для π_H^--критерия Пирсона (см. § 9.3). Тестовая статистика этого критерия записывается в виде

$$\pi_H^- = -2\sum_{i=1}^{N} \ln F(x_i). \tag{9.6.4}$$

Доказывая состоятельность π_N^--критерия для всех альтернатив в классе ω (ω — класс непрерывных распределений, меньших F), Чэпмен [1] отмечает, что асимптотическая нормальность статистики π_N^- при гипотезе сохраняется для всех альтернатив в ω. Поэтому нетрудно вычислить $\overline{\beta}_\pi(\delta)$ и $\beta_\pi(\delta)$. Для этого найдем средние и дисперсии статистики π_N^- при альтернативах G_B и G_H.

$$E_B\{\ln F(x)\} = 1-\delta(1-\ln\delta),$$

$$D_B\{\ln F(x)\} = 1+2\delta^2\ln\delta-(\ln^2\delta)(\delta+\delta^2).$$

$$E_H\{\ln F(x)\} = 1-\delta+u_0\ln\left(1+\frac{\delta}{u_0}\right),$$

откуда

$$\max_{u_0} E_H\{\ln F(x)\} = (1-\delta)[1-\ln(1-\delta)];$$

этот максимум достигается при $u_0 = 1-\delta$. Далее,

$$E_H\{\ln F(x)\}^2 = 2(1-\delta)-u_0[\ln^2(u_0+\delta)-\ln^2 u_0]-$$
$$-2(u_0-\delta)\ln(u_0+\delta)-2u_0\ln u_0.$$

Вычисления показывают, что дисперсия $\ln F(x)$, зависящая от u_0, будет максимальна при $u_0 = 1-\delta$, хотя зависимость от u_0 весьма слаба. При $u_0 = 1-\delta$ дисперсия равна

$$D_H\{\ln F(x)\} = (1-\delta)[2-2\ln(1-\delta)+\ln^2(1-\delta)]-$$
$$-(1-\delta)^2[1-\ln(1-\delta)]^2.$$

Следовательно, при $N\to\infty$

$$\underline{\beta}_\pi(\delta) = \Phi\left\{\frac{z_\alpha+\sqrt{N}[(1-\delta)\ln(1-\delta)+\delta]}{[D_H\{\ln F(x)\}]^{1/2}}\right\}, \tag{9.6.5}$$

$$\overline{\beta}_\pi(\delta) = \Phi\left\{\frac{z_\alpha+\sqrt{N}[\delta(1-\ln\delta)]}{[1+2\delta^2\ln\delta-\ln^2\delta(\delta+\delta^2)]^{1/2}}\right\}. \tag{9.6.6}$$

Аналогичные, хотя и более трудоемкие, вычисления можно проделать и для других критериев. На рис. 9.6.2 приведены для иллюстрации некоторые результаты расчетов, выполненных Чэпменом для пяти тестов при $N=50$ и $\alpha=0{,}05$. По-

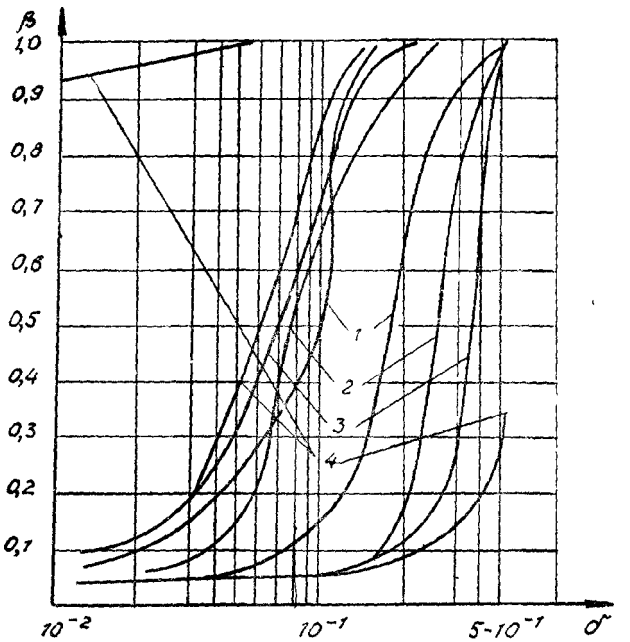

Рис. 9.6.2

скольку верхние и нижние границы для D_N^- -теста оказываются наиболее близкими друг к другу, то Чэпмен делает естественный вывод, что «при отсутствии какой бы то ни было информации о возможных альтернативах D_N^- -тест является наиболее предпочтительным среди рассмотренных, в некотором минимаксном смысле»*.

Если же имеется дополнительная информация о распределениях альтернативы, то всегда можно указать другие более предпочтительные критерии. Например, если известно, что распределения альтернативы в приведенном пространстве имеют вид

$$G^*(u) = u^k,\ k > 1,$$

то π_N^- -критерий является наиболее мощным критерием и максимальная мощность π_N^- -критерия будет гораздо больше максимальной мощности D_N^- -критерия для всех N и δ.

* Возможно, следовало бы сделать более осторожный вывод, поскольку «расстояние» (9.6.1), по которому проводилось сравнение, является базовым для построения именно D_N^--статистики, и может быть именно этим объясняется полученный результат.

Далее, если известно, что альтернативные распределения обладают большими отклонениями от распределения гипотезы в середине области изменения переменной $u(0 \leqslant u \leqslant 1)$, то целесообразнее применять ω_N^2-критерий, нежели Ω_N^2-критерий. В то же время Ω_N^2-критерий более чувствителен к отклонениям на «хвостах» распределений.

В тех случаях, когда альтернативные распределения в канонической форме лежат сначала ниже распределения при гипотезе, а затем симметрично выше, то целесообразнее использовать статистику Купера [1] $V_N = D_N^+ + D_N^-$, нежели просто статистику D_N. В следующем параграфе мы покажем, что именно такой характер имеют приведенные альтернативы для ряда масштабных альтернатив в исходном пространстве. Комо [1] показал, что V_N-тест обладает гораздо большей мощностью нежели D_N-тест для масштабных альтернатив. Исследованиям границ мощности критериев типа Колмогорова-Смирнова занимались Смирнов [4], Бирнбаум [2], Мэсси [1, 2] и другие. Интересным результатом Мэсси [2] является возможность оценивания необходимого объема выборки для достижения D_N-критерием заданной мощности при фиксированном расстоянии $\delta = \sup_x |F - G|$ (см рис. 9.6.2). С помощью этого результата Мэсси показывает, что при больших объемах выборки эти критерии оказываются значительно лучшими, чем χ^2-критерий. Например, на рис. 9.6.2 приведены результаты Мэсси, показывающие, насколько больший объем выборки требуется для χ^2-критерия по сравнению с D_N-критерием для достижения мощности не меньше 0,5 при заданном расстоянии δ. Однако эксперименты Дурбина [1] с умеренными объемами выборки не позволяют утверждать, что D_N-критерий всегда обладает большей мощностью, чем χ^2-критерий.

Аналогичное сравнение ω^2_N- и χ^2-критериев провели Кац, Кифер и Вольфовиц [1]. Они показали, что при достаточно малых δ и $\alpha < 1/2$ для достижения минимальной мощности 0,5 χ^2-критерий требует асимптотически N наблюдений, тогда как ω_N^2-критерию потребуется всего лишь $\alpha N^{4/5}$ наблюдений.

Другой подход к исследованию мощностных свойств тестов, как мы уже отмечали, заключается в рассмотрении асимптотического поведения мощности при стремлении альтернатив к гипотезе с увеличением объема выборки. Альтернативы задаются в виде

$$G(x) = F(x) + \frac{1}{\sqrt{N}} H(x), \tag{9.6.7}$$

где $H(x)$ — фиксированная функция с такими свойствами, что $G(x)$ есть функция распределения. Качества тестов оцениваются по асимптотическому поведению их мощности при $N\to\infty$. Поскольку альтернатива задана, можно найти оптимальный тест для проверки гипотезы против альтернативы (9.6.7) и найти его асимптотическую мощность. Это позволяет определить асимптотическую относительную эффективность (см. § 2.8) рассматриваемого теста по сравнению с оптимальным. К этому классу работ относятся исследования Смирнова [4], Чибисова [1], Вайсса [3], Рамачандрамурти [1] и других. Для примера приведем результаты Чибисова. Оказалось, что при нормальной гипотезе $\Phi(x)$ и альтернативах вида $\Phi(x-\mu)$ и $\Phi\left(\frac{x}{1+\theta}\right)$ относительные асимптотические эффективности (e_μ и e_θ соответственно) ω_N^2-теста по сравнению с оптимальным при альтернативе вида (9.6.7) равны:

при $\alpha=1-\beta=0{,}05$ $e_\mu\geqslant 0{,}53$, $e_\theta\geqslant 0{,}18$;

при $\alpha=1-\beta=0{,}01$ $e_\mu\geqslant 0{,}56$, $e_\theta\geqslant 20$;

при $\alpha=1-\beta=0{,}001$ $e_\mu\geqslant 0{,}58$, $e_\theta\geqslant 0{,}21$.

Эти данные показывают, что ω_N^2-критерий значительно более чувствителен к сдвигу, нежели к масштабу. Определенное представление о мощностных свойствах D_N^--критерия дают результаты Рамачандрамурти, который показал, что питмановская эффективность D_N^--критерия по отношению к оптимальному для нормальных альтернатив сдвига (t-критерию Стьюдента) не опускается ниже чем 0,36.

Третий подход к изучению мощностных свойств тестов согласия основан на асимптотическом представлении тестовых статистик через случайные процессы, которое уже упоминалось в § 9.3. Как и при истинности гипотезы, статистика, вычисленная по выборке из альтернативного распределения, асимптотически эквивалентна некоторому случайному процессу $z(t)$ броуновского типа. Превышение тестовой статистикой критического уровня асимптотически эквивалентно пересечению реализаций $z(t)$ этого процесса некоторой границы $a(t)$, что в принципе и позволяет вычислить асимптотическую мощность теста*.

Продемонстрируем технику этого подхода на альтернативах сдвига и масштаба. При альтернативе $F_\theta(x)$ все статистики структуры d представимы через процесс вида

* Правда, если $a(t)$ не является линейной или кусочно-линейной функцией, пока не найдено достаточно простых методов, позволяющих аналитически вычислить эту вероятность.

$$z_N(t)=\sqrt{N}\{F_N[F_\theta^{-1}(t)]-t\} \tag{9.6 8}$$

(смотри (9.3 10)—(9 3 14). Например, D_N-статистика выразится как

$$D_N=\sup_{0\leqslant t\leqslant 1}|z_N(t)-\sqrt{N}\{F[F_\theta^{-1}(t)-t\}|. \tag{9 6 9}$$

Следовательно, мощность D_N-критерия запишется как

$$\beta_{D_N}(F_\theta)=1-P_0\{|z_N(t)-\sqrt{N}[FF_\theta^{-1}(t)-t]|<C_\alpha(D_N)\},$$

где $C_\alpha(D_N)$ — критическое значение D_N-статистики. Квайд [1] дал следующее обобщение теоремы Донскера [1] — Гихмана и Скорохода [1], которое потребуется при вычислении асимптотической мощности.

Теорема 9.6.1. Пусть $\{a_N(t)\}$ — последовательность вещественных функций, определенных на [0, 1], которая сходится при $N\longrightarrow\infty$ равномерно в [0,1] к некоторой функции $a(t)$. Тогда во всех точках непрерывности справа

$$\lim_{N\to\infty}P\{|z_N(t)-a_N(t)|\leqslant C_\alpha(D_N)\}=$$
$$=P\{|z(t)-a(t)|\leqslant C_\alpha(D_N)\},\quad 0\leqslant t\leqslant 1.$$

Нахождение асимптотической мощности тестов согласия при альтернативах сдвига и масштаба значительно упрощается благодаря результатам Гельцера и Пайка [1]. Для масштабной альтернативы вида $F_\theta(x)=F[(1+\theta)x]$ они доказали следующую теорему:

Теорема 9.6.2. Пусть $F(x)$ имеет плотность $f(x)$, для которой функция $f(F^{-1}(t))$ непрерывна на (0,1) и $\lim_{|x|\to\infty}xf(x)=0$.

Тогда

$$\lim_{\theta\to 0}\theta^{-1}\{F[x(1+\theta)^{-1}]-F(x)\}=-xf(x)$$

равномерно по x для всех $x\in(-\infty,\infty)$.

Аналогично для альтернативы сдвига $F_\theta(x)=F(x-\theta)$ и при условиях теоремы (9.6.2), где условие $\lim_{|x|\to\infty}xf(x)=0$ заменяется на $\lim_{|x|\to\infty}f(x)=0$, легко доказать, что

$$\lim_{\theta\to 0}\theta^{-1}\{F(x+\theta)-F(x)\}=-f(x).$$

Вследствие этих теорем, выбрав $\theta_N=\delta/\sqrt{N}$, $0<\delta<\infty$, имеем для масштабной альтернативы

$$\lim_{N\to\infty}|\sqrt{N}\{F[F_{\theta_N}^{-1}(t)]-t\}+\delta F^{-1}(t)\,f\,[F^{-1}(t)|=0,$$

а для альтернативы сдвига

$$\lim_{N\to\infty}\left|\sqrt{N}\left[F[F_{\theta_N}^{-1}(t)]-t\right]+\delta f[F^{-1}(t)]\right|=0,$$

равномерно по t для всех $t\in[0,1]$.

Применим эти результаты и теорему 9.6.1 для вычисления мощностей конкретных тестов. Например, для D_N-теста при масштабной альтернативе имеем:

$$\lim_{N\to\infty}\beta_{D_N}(F_\theta)=1-P\{|z(t)-s(t,\delta)|<C_\alpha(D_N)\},$$

где

$$s(t,\delta)=-\delta F^{-1}(t)\cdot f[F^{-1}(t)].$$

Аналогично, для D_N^+ при той же альтернативе получаем

$$\lim_{N\to\infty}\beta_{D_N^+}(F_\theta)=1-P\{z(t)\leqslant s(t,\delta)+C_\alpha(D_N^+)\},\quad 0\leqslant t\leqslant 1.$$

Для семейства симметричных относительно нуля распределений и масштабных альтернатив Гельцер и Пайк [1] предложили тест, основанный на абсолютных значениях наблюдений. Тестовая статистика записывается в виде

$$D_N^+=\sup_{0\leqslant v<\infty}\{\sqrt{N}\,[F_N^*(v)-F^*(v)]\},$$

где $V=|X|$ и $F^*(x)=2F(v)-1$. Для этого теста

$$\lim_{M\to\infty}\beta_{D_N^*}(F_\theta)=1-P\{z(t)<s^*(t,\delta)+C_\alpha(D_N^*)\},$$

где

$$s^*(t,\delta)=-\delta F^{*-1}(t)\,f^*[F^{*-1}(t)]=$$
$$=-\delta F^{-1}\left(\frac{t+1}{2}\right)\cdot 2f\left[F^{-1}\left(\frac{t+1}{2}\right)\right].$$

Вычисление мощностей в случае, когда функции $s(t,\delta)$ или $s^*(t,\delta)$ являются линейными, производится сравнительно просто; Я. Гаек и З. Шидак [1] приводят многочисленные примеры такого рода. Некоторые результаты можно получить, если нелинейные функции s и s^* ограничить кусочно линейными функциями (смотри, например, Квайд [1], Гельцер [1]), хотя уже и в этом случае возникают большие трудности. Если нас интересует лишь качественное сравнение мощностей, то можно воспользоваться тем фактом, что для неравенства $\beta_1>\beta_2$ достаточным условием является неравенство $s_1(t,\delta)<>s_2(t,\delta)$. Например, для двойного экспоненциального распределения и масштабной альтернативы $s^*(t,\delta)\leqslant s(t,\delta)$ для

всех t и δ. Следовательно $D_N^{\wedge}$-тест в этом случае обладает большей мощностью, чем D_N^{+}-тест.

К сожалению, нелинейность границ $a(t)$, невыполнение неравенств между ними (в случае их пересечения), а иногда — невозможность получения $a(t)$ в явном виде, существенно ограничивают возможности данного метода.

Таким образом, рассмотрев известные аналитические подходы к сравнению тестов по их мощности, мы приходим к общему утверждению, что эти подходы позволяют сравнивать не любые тесты и не при любых альтернативах.

§ 9.7. ОБ ОДНОМ ЭВРИСТИЧЕСКОМ МЕТОДЕ СРАВНЕНИЯ МОЩНОСТЕЙ ТЕСТОВ СОГЛАСИЯ СТРУКТУРЫ

Можно сформулировать несколько типичных задач, в которых возникает необходимость сравнения тестов по их мощности.

1. Если задано несколько альтернатив G_1, ..., G_k, то как упорядочатся мощности $\beta_S(F, G_1)$, ..., $\beta_S(F, G_k)$ теста согласия с F, основанного на заданном расстоянии ρ (заданной статистике S)?

2. Если задана конкретная альтернатива G и некоторый класс тестов на однотипных расстояниях*, то такой тест из этого класса будет наиболее мощным при данной альтернативе?

3. Если задан набор альтернатив G_1, ..., G_k и набор тестов S_1, ..., S_l, основанных на разнотипных расстояниях, то какие пары «тест — альтернатива» предпочтительны в смысле мощности?

4. Если задана альтернатива G и два разнотипных теста, то какой из них предпочтителен?

Если бы у нас были методы аналитического вычисления мощностей любых тестов при любых альтернативах, ответить на эти вопросы было бы просто. Однако сложность (а иногда невозможность) вычисления мощностей вынуждают нас искать других путей. На то, что такие пути должны существовать, указывает сам характер перечисленных задач: нам не требуется знать ни точного значения мощностей, ни даже только их разности; мы вполне удовлетворились бы знаком разности. Таким образом, мы приходим к следующей задаче: для каких тестов $\{S\}$ и при каких альтернативах $\{\tau(u)\}$ можно сделать суждения типа $\beta_S(\tau_1) \geqslant \beta_S(\tau_2)$ или $\beta_{S_1}(\tau) \geqslant \beta_{S_2}(\tau)$ без вычисления мощностей (или распределений статистики)?

* Например, расстояния d_4 (§ 9.2) при разных Ψ однотипны, а d_2 и d_4 — разнотипны.

По отношению к тестам согласия структуры d в ряде случаев на этот вопрос можно ответить, опираясь на некоторые общие свойства классов тестов, рассматриваемых альтернатив и тестовых статистик. Перейдем к обсуждению этих свойств*.

Согласно теореме Чэпмена (см. § 9.6) о мощности монотонного теста структуры d при альтернативах, подчиняющихся неравенству, можно сделать суждение сразу: для таких тестов, если $\tau_1(u) \geqslant \tau_2(u)$, $u \in [0,1]$, $\beta(\tau_1) \geqslant \beta(\tau_2)$. В этом случае сравнение мощностей сводится к сравнению альтернатив.

Для удобства будем рассматривать альтернативы в канонической форме, т. е. $\tau(u) = GF^{-1}(u)$, $0 \leqslant u \leqslant 1$. В таком представлении любая гипотеза F соответствует равномерному в $[0,1]$ распределению $\tau_0 = u$; в силу монотонности F сдвиговые альтернативы сохранят односторонность (т. е. либо $\tau \geqslant \tau_0$, либо $\tau \leqslant \tau_0$). При фиксации медианы и квантиля уровня, отличного от 0,5 (например, 0,95), все симметричные распределения упорядочатся по свойству, которое Гаек [1] предложил назвать «степенью затянутости хвостов» (например, некоторые стандартные распределения располагаются следующим образом в порядке возрастания степени затянутости хвостов: нормальное, логистическое, двойное экспоненциальное (распределение Лапласа), распределение Коши).

Пример 9.7.1. Оказывается, что степень затянутости хвостов влияет на различие между гипотезой и альтернативой в канонической форме: на рис. Д.5.1 приведены графики альтернатив правостороннего сдвига при одинаковом ($\theta = 0{,}4$) сдвиге; нумерация кривых соответствует номеру распределения в вышеприведенном ряду. Опираясь на частичную упорядоченность тестов структуры d и характер указанных альтернатив сдвига (рис. Д.5.1), можно сразу сделать вывод, что поскольку для всех u $\tau_1 \geqslant \tau_2 \geqslant \tau_3$, $\beta(\tau_1) \leqslant \beta_s(\tau_2) \leqslant \beta_s(\tau_3)$. Суждения о $\beta_s(\tau_4)$ сделать нельзя, поскольку неравенство $\tau_4 \leqslant \tau_1$ выполняется не для всех u.

Рассмотренный пример ставит два вопроса: 1) существуют ли другие ситуации (а не только гипотеза сдвига), в которых можно делать столь же однозначные выводы о сравнении мощностей? 2) какие заключения можно сделать в тех случаях, когда неравенство между альтернативами выполняется не для всех u?

Ответ на первый вопрос можно получить, лишь обобщив понятие частичной упорядоченности, введенное только для од-

---

* Отметим, что попытка построения метода сравнения мощностей без их вычисления изложена в работе Тарасенко и Шуленина [2]. Однако Д. М. Чибисов [3] указал на недостаточную строгость обоснования метода. В данном параграфе метод изложен с учетом этой критики.

носторонних альтернатив. Такое обобщение становится очевидным, как только мы объединим воедино условия упорядоченности для правосторонней ($\tau_1 \geqslant \tau_2$) и для левосторонней альтернатив ($\tau_2 \geqslant \tau_1$): если для любых $u \subset [0,1]$ выполняется условие

$$|\tau_1(u) - \tau_0(u)| \leqslant |\tau_2(u) - \tau_0(u)|, \qquad (9.7.1)$$

то

$$\beta_s(\tau_1) \leqslant \beta_s(\tau_2).$$

В такой форме понятие частичной упорядоченности тестов структуры становится применимым и к масштабным альтернативам. (Аналогично обобщается понятие монотонности)

Пример 9.7.2. Затянутость хвостов влияет на степень удаленности альтернативы от гипотезы не только при сдвиге, но и при изменении масштаба. На рис. Д.5.2 приведены графики альтернатив масштаба (для $\sigma = 1{,}25$; $\sigma_0 = 1$), номера кривых соответствуют нумерации примера 9.7.1. Легко видеть, что условие частичной упорядоченности выполняется для всех приведенных масштабных альтернатив, и мы можем в этом случае сделать вывод, что для монотонных тестов структуры d $\beta_s(\tau_1) \geqslant \beta_s(\tau_2) \geqslant \beta_s(\tau_3) \geqslant \beta_s(\tau_4)$.

Для альтернативы, каноническое представление которой τ^* имеет сложную форму, суждение о мощностных свойствах частично упорядоченного теста можно сделать, если можно подобрать такую альтернативу $\tau(u, \alpha)$, которая, с одной стороны, позволяла бы достаточно просто вычислить мощность, а с другой — характеризовалась бы некоторым параметром α, меняя который, мы могли бы заключить τ^* в полосу: $|\tau(u, \alpha_1) - \tau_0| \geqslant |\tau^*(u) - \tau_0| \geqslant |\tau(u, \alpha_2) - \tau_0|$. Тогда снова $\beta_s(\tau(u, \alpha_1)) \geqslant \beta_s(\tau^*) \geqslant \beta(\tau(u, \alpha_2))$.

Обратимся теперь ко второму вопросу. Суждения, получаемые подобно тому, как это сделано в примерах 9.7.1 и 9.7.2, справедливы при любых объемах выборки, но эта общность достигается ценой строгого выполнения неравенств (9.7.1) для всех u. Именно поэтому мы ничего не могли сказать о мощности при τ_4 в примере 9.7.1. Если теперь учесть некоторые свойства тестовых статистик, то можно продвинуться дальше. Вспомним, что статистика есть оценка некоторого «расстояния» $\rho(F, G)$, и что, как правило, эта оценка является состоятельной. В силу этого для тестов, основанных на таких статистиках, можно утверждать, что если $\rho(\tau_1, \tau_0) > \rho(\tau_2, \tau_0)$, то найдется такое N_0 (объем выборки), что при $N > N_0$ $\beta_\rho(\tau_1) > \beta_\rho(\tau_2)$.

Пример 9.7.3. В условиях примера 9.7.1 при достаточно большом N мощность теста Смирнова при τ_4 превысит мощности при остальных трех альтернативах.

Недостатком суждений, подобных только что приведенному, является неопределенность величины N_0, однако это не лишает их полезности.

Ясно, что мощность теста неоднозначно определяется величиной «расстояния», однако ясно также, что верхняя грань мощности монотонно возрастает с ростом ρ. Это позволяет делать некоторые суждения о мощностных свойствах тестов путем сравнения «расстояний».

Пример 9.7.4. Рассмотрим класс тестов, порождаемых расстоянием вида $\rho=\int\varphi(F, G)\ \psi(F)\,dF$, где φ — заданная непрерывная функция аргументов, а ψ — весовая функция. Пусть F и G заданы. Спрашивается, при каком $\psi(F)$ верхняя грань мощности теста будет наибольшей? Ответ сводится к нахождению $\psi(F)$, обеспечивающего супремум ρ при заданных F, G и φ. Выразим ρ через канонически приведенные F и G: $\rho=\int\varphi(u, \tau(u))\ \psi(u)\,du$. Супремум ρ достигается при $\psi(u)=\sum\limits_k\delta(u-u_k)$, где $\{u_k\}$ — значения u, при которых $\varphi(u, \tau)$ имеет максимумы. Таким образом, тест из рассматриваемого класса, имеющий максимальную огибающую мощности, есть тест, порождаемый расстоянием $\rho=\sum\limits_k\varphi(u_k, \tau(u_k))$. Это позволяет утверждать, например, что для масштабных альтернатив, $\tau(u)$ которых один раз пересекает $\tau_0(u)$, V_N-тест Купера ($V_N=D_N^+ + D_N^-$) будет более мощным, чем D_N-тест Колмогорова.

Пример 9.7.5. Покажем, что при любых симметричных распределениях в случае масштабной альтернативы $G(x)=F(\sigma x)$ тест Смирнова на абсолютных величинах выборочных значений будет иметь всегда значительно большую верхнюю грань мощности, чем тесты Смирнова и Колмогорова на исходных выборочных значениях, и близкую (возможно, равную) мощность с тестом Купера. Для этого сравним соответствующие «расстояния» $\rho_i=\sup(\tau_i(u)-u)$, $i=1, 2$, где τ_1 — приведенная альтернатива для исходной выборки, а τ_2 — приведенная альтернатива для абсолютных значений. При симметричности F и G имеем:

$$\tau_1(u)=F(\sigma F^{-1}(u)),\ \tau_2(u)=2F\left(\sigma F^{-1}\left(\frac{u+1}{2}\right)\right)-1. \quad (9.7.2)$$

Нахождение максимумов функций $\rho_i=\tau_i(u)-u$, $i=1, 2$ приводит к следующим уравнениям для значений u_{mi}, соответствующих положениям этих максимумов:

$$\sigma f[\sigma F^{-1}(u_{m1})]=f(F^{-1}(u_{m1})),$$
$$\sigma f\left[\sigma F^{-1}\left(\frac{u_{m2}+1}{2}\right)\right]=f\left[F^{-1}\left(\frac{u_{m2}+1}{2}\right)\right]. \quad (9.7.3)$$

Легко видеть, что $u_{m1}=\frac{u_{m2}+1}{2}=u_0$. Отсюда

$$\rho_2=2\mathrm{F}[\sigma F^{-1}(u_0)]-2u_0=2\{F[\sigma F^{-1}(u_0)]-u_0\}=2\rho_1 \qquad (9\,7\,4)$$

Столь большое увеличение расстояния типа Колмогорова — Смирнова между гипотезой и альтернативой при переходе от выборочных значений к их абсолютным величинам и является основанием для утверждений, сделанных в первом предложении данного примера. Суждение о тесте Купера следует из того, что $\tau_1(u)-u$ имеет два экстремума, каждый из которых равен ρ_1 (см. рис. 9.7.1).

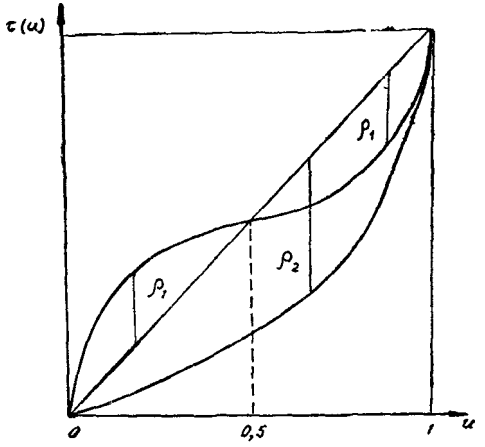

Рис 9 7 1

Как видим, предлагаемая методика позволяет делать выводы о мощностях тестов без их вычисления Следующий параграф содержит экспериментальные данные, подтверждающие эти выводы.

§ 9.8 ОБ ЭКСПЕРИМЕНТАЛЬНОМ СРАВНЕНИИ КРИТЕРИЕВ СОГЛАСИЯ СТРУКТУРЫ *d*

Из предыдущих параграфов становится ясным, что теоретическое исследование мощностных свойств тестов и эвристические способы не всегда могут в полной мере удовлетворить потребность в указании предпочтительных критериев согласия при некоторых заданных условиях. Современная вычислительная техника предоставляет выход из этого положения, которым, судя по литературе, все чаще начинают пользоваться исследователи. Этот выход состоит в экспериментальном сравнении тестов.

Техника статистического эксперимента в принципе достаточно проста. Сначала генерируется выборка из любого ин-

тересующего нас конкретного распределения Затем по этой выборке вычисляются сравниваемые тестовые статистики Полученные значения сравниваются с критическими значениями для соответствующих статистик (при одинаковых уровнях значимости). Проводя такой эксперимент многократно на различных выборках одинакового объема, можно вычислить относительные частоты принятия альтернативы различными тестами. Эти частоты и являются оценками соответствующих мощностей, и их сравнение позволит сделать суждения о предпочтительности того или иного теста в данной ситуации.

Некоторая трудность может возникнуть из-за того, что точные или асимптотические распределения тестовых статистик при нулевой гипотезе известны не для всех статистик В таких случаях естественно воспользоваться статистическими оценками критических значений, в качестве которых могут служить соответствующие порядковые статистики реализовавшихся значений тестовых статистик.

Ценность таких статистических экспериментов в значительной мере зависит от того, насколько общие выводы можно сделать на их основе, т. е. от того, в какой мере результаты экспериментирования с фиксированными альтернативами останутся справедливыми для некоторого множества других альтернатив Поэтому весьма важным было бы установление некоторых классов альтернатив, обладающих определенными общими свойствами по отношению к данному классу тестов. Здесь может быть полезной классификация распределений по симметричности, одновершинности и затянутости хвостов, а также стандартизованное (каноническое) представление пар конкурирующих гипотез (см. Добавление к Части I) Статистический эксперимент, который мы опишем ниже, был приведен для ряда критериев согласия и указанных альтер-

Таблица 981

Нормальная альтернативная сдвига

a	T_N^-	π_N^-	ω_N^2	Ω_N^2	D_N^-	$R_\Lambda^-\left(\frac{1}{4}, 1\right)$
0,00	0,01	0,02	0,00	0,00	0,03	0,05
0,10	0,13(0,17)	0,16(0,11)	0,08(0,11)	0,09(0,10)	0,14(0,18)	0,10(0,15)
0,20	0,45(0,46)	0,40(0,40)	0,33(0,29)	0,36(0,29)	0,39(0,39)	0,34(0,32)
0,25	0,53	0,48	0,40	0,40	0,51	0,42
0,30	0,70	0,65	0,54	0,57	0,60	0,57
0,40	0,88(0,91)	0,82(0,77)	0,80(84)	0,83(0,86)	0,80(0,82)	0,73(0,74)
0,50	0,99	0,95	0,94	0,95	0,96	0,86
0,60	1,00	1,00	0,99	1,00	0,99	0,92

Здесь в скобках приведены некоторые результаты повторного эксперимента при других выборках.

Таблица 9.8.2

Альтернатива сдвига Коши

a	T_N^-	π_N^-	ω_N^2	Ω_N^2	D_N^-	$R_N\left(\frac{1}{4}, 1\right)$	$R_N\left(\frac{1}{2}, 1\right)$
0,05	0,19	0,10	0,12	0,14	0,19	0,10	0,20
0,10	0,49	0,25	0,32	0,38	0,50	0,20	0,55
0,15	0,72	0,48	0,62	0,67	0,76	0,35	0,82
0,20	0,89	0,69	0,83	0,85	0,91	0,55	0,93
0,25	0,99	0,86	0,95	0,96	0,99	0,75	0,99
0,30	1,00	0,99	1,00	1,00	1,00	0,91	1,00

натив сдвига и масштаба (Жирова, Тарасенко, Шуленин [1]) при $N=50$, $\alpha=0{,}05$ — табл. 9.8.1—9.8.5.

Зависимости мощности критерия Реньи от параметра b приведены на рис. 9.8.1 для нормального распределения и на рис. 9.8.2 для распределения Коши.

Таблица 9.8.3

Логистическая масштабная альтернатива

θ	D_N	$R_N\left(\frac{1}{2}, 1\right)$	ω_N^2	Ω_N^2	$V_N=D_N^+ + D_N^-$
1,00	0,02	0,03	0,05	0,04	0,02
1,30	0,05	0,03	0,06	0,05	0,16
1,50	0,10	0,06	0,11	0,15	0,63
1,60	0,23	0,03	0,17	0,27	0,70
1,80	0,40	0,10	0,41	0,59	0,92
2,00	0,60	0,22	0,78	0,88	0,98
2,20	0,83	0,33	0,93	0,96	1,00
2,40	0,90	0,46	0,97	0,99	1,00

Таблица 9.8.4

Лапласовская масштабная альтернатива

θ	D_N	$R_N\left(\frac{1}{2}, 1\right)$	ω_N^2	Ω_N^2	$V_N=D_N^+ + D_N^-$
1,00	0,03	0,02	0,03	0,03	0,04
1,30	0,02	0,04	0,04	0,03	0,05
1,50	0,10	0,05	0,07	0,06	0,42
1,60	0,12	0,06	0,05	0,09	0,51
1,80	0,26	0,18	0,30	0,43	0,88
2,00	0,44	0,17	0,58	0,70	0,91
2,20	0,57	0,27	0,75	0,86	0,97
2,40	0,73	0,32	0,88	0,93	1,00
2,60	0,84	0,43	0,95	0,98	1,00

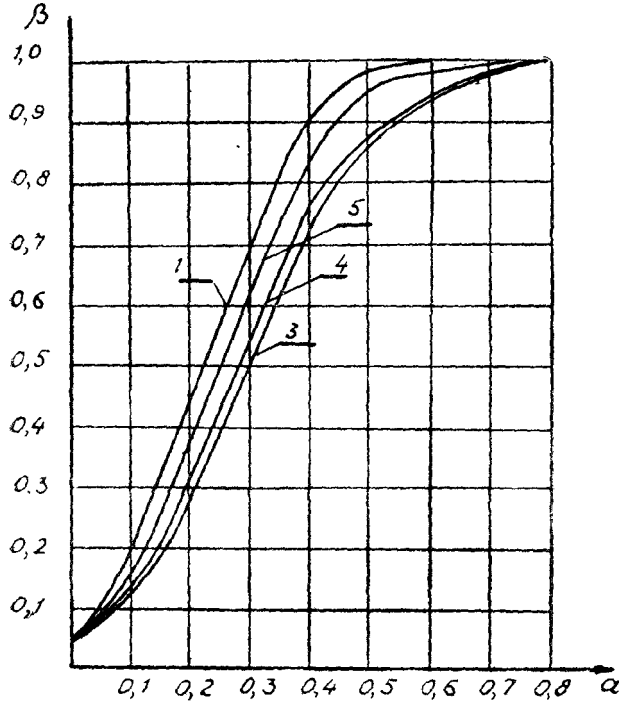

Рис. 9.8.1

Таблица 9.8.5

Масштабная альтернатива Коши

θ	D_N	$R_N\left(\frac{1}{2},1\right)$	ω_N^2	Ω_N^2	V_N
1,00	0,04	0,05	0,03	0,03	0,02
1,30	0,07	0,09	0,08	0,08	0,16
1,50	0,06	0,07	0,04	0,05	0,26
1,60	0,12	0,09	0,09	0,07	0,40
1,80	0,16	0,07	0,12	0,11	0,60
2,00	0,35	0,17	0,25	0,27	0,75
2,20	0,36	0,23	0,40	0,47	0,82
2,40	0,49	0,25	0,57	0,63	0,91
2,60	0,61	0,31	0,69	0,74	0,94

На основании экспериментальных данных можно сделать следующие выводы.

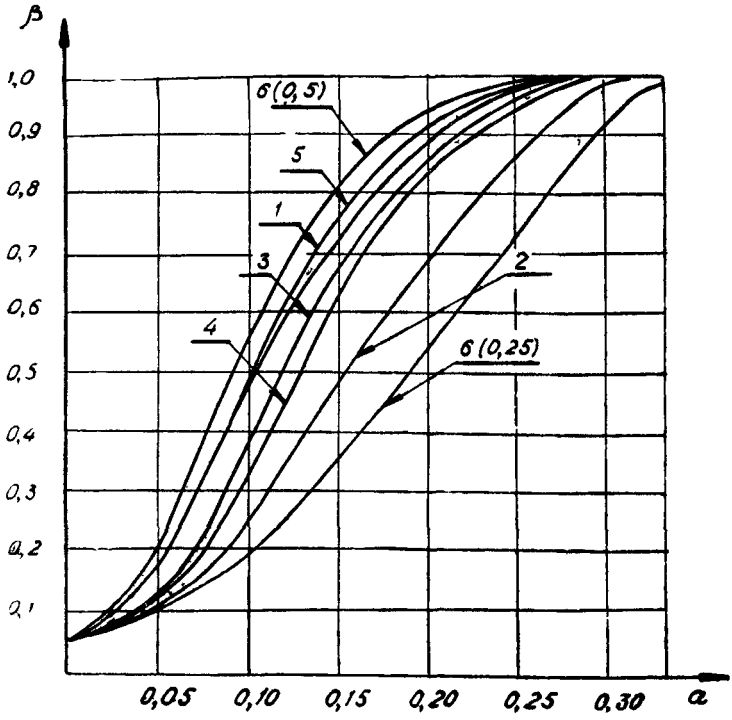

Рис 982

1. Мощности рассмотренных критериев согласия для альтернатив сдвига примера 9.7.1 возрастают при переходе к распределениям с более затянутыми хвостами. Этот факт подтверждает эвристические выводы § 9. 7.

2. Предпочтительность T_N^--теста Мозеса при логической альтернативе сдвига полностью согласуется с результатами работы Белла и Доксума [1], в которой показана локальная оптимальность T_N^--теста при этой альтернативе

3 Мощность R_N^- $(b, 1)$-теста Реньи существенно зависит от параметра b. Для рассмотренных альтернатив сдвига она максимальна при $b=0{,}5$

4 D_N^--тест Смирнова предпочтительней R_N^- $(b, 1)$-теста Реньи для альтернатив сдвига с быстро спадающими хвостами; для распределений с сильно затянутыми хвостами (распределение Коши) картина обратная.

5 Для альтернатив сдвига с быстро спадающими хвостами ω_N^2-тест Крамера-Мизеса-Смирнова предпочтительнее

Ω_N^2-теста Андерсона-Дарлинга; в случае сильно затянутых хвостов результат обратный.

6 Мощности рассмотренных критериев согласия для масштабных альтернатив примера 9 7 1 уменьшаются при переходе от распределений с быстро спадающими хвостами к распределениям с сильно затянутыми хвостами. Этот факт хорошо подтверждает эвристические выводы § 9. 7

7 Для всех рассмотренных типов распределений при масштабных альтернативах V_N-тест Купера имеет большую мощность по сравнению с другими тестами согласия

Приведем теперь некоторые экспериментальные результаты, подтверждающие гипотезу (Тарасенко [7]) о том, что мощность некоторого класса критериев согласия однозначно определяется энтропией альтернативы в канонической форме, совпадающей с информационным расхождением Кульбака I (2; 1)

$$H=-\int_0^1 g^*(u)\ln g^*(u)\,du,$$

где

$$g^*(u)=g(F^{-1}(u))/f(F^{-1}(u)),\quad 0\leqslant u\leqslant 1,$$

а g и f — плотности альтернативного и гипотического распределений

Таблица 986

Результаты эксперимента приведены для масштабных альтернатив

H	Тип распределения	0	D_N	$R_N\left(\frac{1}{2},1\right)$	ω_N^2	Ω_N^2	V_N
	Нормальное	1,65	0,22	0,12	0,25	0,30	0,72
—0,115	Лапласа	1,95	0,21	0,12	0,24	0,32	0,78
	Коши	2,40	0,25	0,15	0,25	0,29	0,75

Эти экспериментальные результаты наводят на мысль о том, что мощности этих критериев согласия несущественно зависят от типа распределения, а определяются лишь энтропией канонической альтернативы. К сожалению, пока не найдено строгого доказательства этой гипотезы, которая в случае справедливости существенно бы упростила задачу нахождения мощностей тестов при различных альтернативах. Для решения этой задачи понадобилось бы лишь знание мощности теста при некоторой простой, позволяющей провести вычисления альтернативе и энтропии любой другой канонической альтернативы

§ 9.9 ТЕОРЕТИЧЕСКОЕ СРАВНЕНИЕ МОЩНОСТНЫХ СВОЙСТВ КРИТЕРИЕВ СОГЛАСИЯ НА ВЫБОРОЧНЫХ ИНТЕРВАЛАХ

Как уже отмечалось (см. § 9.4), обширный класс критериев согласия основан на свойствах выборочных интервалов приведенной выборки (тесты структуры D). В данном параграфе мы займемся рассмотрением мощностных свойств этого класса тестов.

Трудности нахождения распределений (точных и асимптотических) статистик на выборочных интервалах связаны, во-первых, с зависимостью интервалов (поскольку $\sum_{i=1}^{N+1} \Delta_i \equiv 1$), и, во-вторых, с тем, что некоторые статистики этого класса не являются асимптотически нормальными.

Эти трудности и объясняют немногочисленность и недостаточную общность результатов теоретического исследования мощности тестов структуры D. Приведем полученные к настоящему времени теоретические данные по этому вопросу.

Вайсс [5] доказал следующую теорему.

Теорема 9.9.1. Если плотность $g(x)$ приведенной альтернативы имеет конечное число разрывов непрерывности и $0<A\leqslant g(x)\leqslant B<\infty$ для всех x, то статистики Кимбалла $T_1^{\alpha}=\sum_{i=1}^{N+1}\Delta_i^{\alpha}$ асимптотически нормальны*: распределение величины

$$\frac{N^{\alpha-\frac{1}{2}}T_1^{\alpha}-\sqrt{N}\,\Gamma(\alpha+1)\int_0^1 g^{1-\alpha}\,dx}{\sqrt{\Gamma(2\alpha+1)-2\alpha\Gamma^2(\alpha+1)\int_0^1 g^{1-2\alpha}\,dx-\left[(\alpha-1)\,\Gamma(\alpha+1)\int_0^1 g^{1-\alpha}\,dx\right]^2}} \tag{9.9.1}$$

стремится при $N\to\infty$ к стандартному нормальному распределению.

Как видим, даже для сравнительно простого класса статистик асимптотическое распределение удается получить лишь при некоторых ограничениях на тип альтернатив. Поэтому поиски продолжаются. Один из путей обхода трудностей, связанных с зависимостью интервалов, состоит в использовании асимптотически эквивалентного их представления

* Строгое доказательство этой теоремы было дано Вайссом для канонических плотностей ступенчатой формы; в этой же работе он эвристически распространил полученный результат на непрерывные плотности. Пайк [1] советует с осторожностью относиться к такому обобщению.

через независимые и одинаково (экспоненциально) распределенные случайные величины (см. § 4.4). Но даже и на этом пути основные результаты получены лишь при рассмотрении последовательности альтернатив, определенным образом сходящейся к гипотезе.

Определим альтернативную плотность в приведенном пространстве как

$$g(x)=1+\frac{l(x)}{N^{\delta}}, \tag{9.9.2}$$

где $l(x)$ — фиксированная функция, обладающая свойствами

$$\int_0^1 l(x)dx=0, \int_0^1 l^2(x)dx=M<\infty, \tag{9.9.3}$$

а $\delta>0$ — постоянная, характеризующая скорость сходимости альтернативы к гипотезе. Для такой альтернативы Вайсс [6] доказал следующее утверждение.

Теорема 9.9.2. Для любой измеримой в N-мерном пространстве области R_N

$$\lim_{N\to\infty}\left|\int_{R_N} f(v_1,\ldots, v_N)dv_1\ldots dv_N-\int_{R_N} q(v_1\ldots, v_N)dv_1\ldots dv_N\right|=0, \tag{9.9.4}$$

где $f(\Delta_1, \ldots, \Delta_N)$ — совместная плотность вероятностей выборочных интервалов при распределении (9.9.2), а $q(z_1,\ldots,z_N)$ — совместная плотность случайных величин $\{z_i\}$, $z_i=W_i/T_N'$, где

$$W_i=\frac{Y_i}{g\left[G^{-1}\left(\frac{i}{N+1}\right)\right]}, \quad T_N'=\sum_{i=1}^{N+1} W_i, \quad i=1,\ldots, N,$$

а $Y_1,\ldots, Y_{N+1}$ — независимые экспоненциально распределенные случайные числа с единичным средним.

С помощью теоремы 9.9.2 Вайсс [6] показал, что для статистики T_1^2 ((9.4.5) при $\alpha=2$), «экспоненциальной копией» которой является выражение

$$T_1^2=\frac{1}{(T_N')^2}\sum_{i=1}^{N+1}\left\{\frac{Y_i}{g\left[G^{-1}\left(\frac{i}{N+1}\right)\right]}\right\}^2, \tag{9.9.5}$$

асимптотическая мощность T_1^2 — теста при альтернативе (9.9.2) может быть выражена в виде

$$\beta_{T_1^2}=1-\Phi\left[k_{\alpha^*}-N^{\frac{1}{2}-2\delta}\int_0^1 l^2(x)\,dx\right], \tag{9.9.6}$$

где k_{α^*} — корень уравнения $1-\Phi(k_{\alpha^*})=\alpha^*$, α^* — уровень значимости. Геберт [1] обобщил этот результат Вайсса на случай произвольного α:

$$\beta_{T_1^2}=1-\Phi\left[k_{\alpha^*}-N^{\frac{1}{2}-2\alpha}\cdot r(\alpha)\cdot M\right], \tag{9.9.7}$$

где

$$r(\alpha)=\frac{\alpha(\alpha-1)\Gamma(\alpha+1)}{2\sqrt{\Gamma(2\alpha+1)-(\alpha^2+1)\Gamma^2(\alpha+1)}}. \tag{9.9.7}$$

Максимизируя (9.9.7) по α, Геберт показал, что наибольшая мощность при альтернативе вида (9.9.2) достигается при $\alpha=2$ (T_1^2 — тест Гринвуда).

Из (9.9.6) и (9.9.7) видно, что при $N\to\infty$ $\beta_{T_1^\alpha}$ не будет стремиться к α^* или 1 лишь при $\delta=1/4$. Этот факт позволяет дать правильное толкование результату Чибисова [2], который показал, что асимптотическая относительная эффективность тестов структуры D по отношению к тестам структуры d равна нулю при альтернативе (9.6.2) и $\delta=\frac{1}{2}$. То, что тесты структуры d для таких альтернатив оказываются существенно лучше, чем тесты структуры D, еще не означает предпочтительности первых во всех случаях. Вайсс [7] привел пример (экспоненциальная альтернатива правостороннего сдвига, который связан с уровнем значимости, и тест основан на максимальном выборочном интервале), в котором картина становится полностью противоположной.

Пользуясь «экспоненциальными копиями» тестовых статистик, Блюменталь [4] сравнил мощности тестов Гринвуда и Дарлинга для альтернативы (9.9.2) при $\delta=\frac{1}{4}$. Для теста Гринвуда (T_1^2) асимптотическая мощность есть (9.9.6), а для теста Дарлинга ($S=\sum\ln\Delta_i$) она равна $1-\Phi\left[k_\alpha-\int t^2dx\cdot\left(\frac{2\pi}{\sqrt{6}}-1\right)^{-1}\right]$. Таким образом, питмановская эффективность теста Дарлинга по отношению к тесту Гринвуда равна $\left(\frac{2\pi}{\sqrt{6}}-1\right)^{-2}\approx 0{,}47$.

Для альтернатив $g(x)$, график которых состоит из конечного числа горизонтальных отрезков («ступенчатая» плотность), Вайсс также получил ряд результатов. Например, он показал (Вайсс [2]), что совместное асимптотическое распределение $\Delta_{(1)}$ и $\Delta_{(N+1)}$ равно

$$\lim_{N\to\infty} P_r\left[\Delta_{(1)} > \frac{u}{(N+1)^2},\ \Delta_{(N+1)} < \frac{\log(N+1)+\log M-\log v}{M(N+1)}\right] = \exp\left[-Su+Bv\right], \qquad (9.9.8)$$

где $0<u<v<1$, $M=\min\limits_x g(x)>0$, $S=\int\limits_0^1 g^2dx$, B — сумма интервалов, на которых $g(x)=M$. Это распределение позволяет найти асимптотическую мощность тестов на минимальном и/или максимальном интервалах.

Сетураман и Рао [1], пользуясь асимптотически эквивалентным представлением статистик в виде случайного процесса (см. § 9.6), нашли питмановские константы эффективности для статистик T_1^a ($\alpha=0(0,5)4$), T_2 и T_3 (см. § 9.4), при альтернативе (9.9.2), $\delta=1/4$. По их результатам, $T_1{}^2$-тест имеет наибольшую (равную 1) эффективность среди всех T_1^a-тестов, T_2-тест имеет эффективность 0,5726, а T_3-тест — 0,3876*.

Некоторые возможности в изучении мощностных свойств тестов структуры D представляет рассмотрение канонической альтернативы, являющейся равномерным в $[\delta, 1]$ распределением, $\delta\in(0,1)$. Интересно, что такая каноническая альтернатива соответствует правостороннему сдвигу экспоненциального распределения в исходном пространстве, т. е. не является только искусственным теоретическим подспорьем. Преимущество данной альтернативы состоит также в том, что она позволяет для некоторых тестов воспользоваться результатами, полученными при истинности гипотезы, так как при альтернативе сохраняется равномерность распределения, и остается учесть лишь особенности, связанные с появлением одного «большого» интервала $\delta+\Delta_1$.

Представим статистику $S=\sum\limits_{i=1}^{N+1} h(\Delta_i)$ в эквивалентной форме $S=S^1+h(\delta+\Delta_1)$, где $S'=\sum\limits_{i=2}^{N+1} h(\Delta_i)$. Очевидно, что при $N\to\infty$ $S\to S_\delta+h(\delta)$, где S_δ — статистика S в предположении, что интервалы $\{\Delta_i\}$ соответствуют равномерному в $[\delta, 1]$ распределению. Дарлинг [2] предложил метод для получения асимптотического распределения статистики при гипотезе (см. § 9.4). В силу вышесказанного этот метод можно распространить на вычисление асимптотического распределения при данной альтернативе, а следовательно, по-

* Отметим расхождение последней цифры с результатом Блюменталя [4]; мы не смогли найти причину расхождения, однако есть основания полагать, что данные Блюменталя верны.

лучить асимптотическую мощность S-теста или оценку его мощности при конечных N, допускающих аппроксимацию действительного распределения асимптотическим.

Таблица 9.9.1

T_i	δ	Среднее	Дисперсия
T_1^2	$\delta=0$	$\frac{2}{N}$	$\frac{4}{N^3}$
	$\delta>0$	$\delta^2+\frac{2}{N-1}(1-\delta)^2$	$\frac{4}{(N-1)^3}(1-\delta)^4$
T_2	$\delta=0$	$1-C-\ln(N+1)$	$\frac{1}{N+1}\left(\frac{\pi^3}{3}-1\right)$
	$\delta>0$	$(1-\delta)(1-C-\ln N)+$ $+(1-\delta)^2\ln(1-\delta)+\delta\ln\delta$	$\frac{1-\delta}{N}\left(\frac{\pi^2}{3}-1\right)$
T_5	$\delta=0$	$-(N+1)(\ln N-C)$	$(N+1)\left(\frac{\pi^2}{6}-1\right)$
	$\delta>0$	$(N+1)\ln(1-\delta)-$ $-N[\ln(N-1)+C]$	$N\left(\frac{\pi^2}{6}-1\right)$

Для примера приведем полученные таким образом результаты для T_1^2- и T_5-тестов, функции $h(x)$ которых удовлетворяют отмеченному условию. При гипотезе и при рассматриваемой альтернативе тестовые статистики являются асимптотически нормальными. Параметры соответствующих распределений приведены в таблице 9.9.1. При $\delta>0$ формулы получены соответствующим обобщением формул (9.4.6) и подобных формул для T_5. Выражения для функций мощности получаются в виде:

$$\beta_{T_1^2}(\delta)=1-\Phi\left[\frac{z_\alpha}{(1-\delta)^2}+\left(\frac{\delta}{1-\delta}\right)^2\left(N^{\frac{1}{2}}-\frac{1}{2}N^{\frac{3}{2}}\right)\right]; \quad (9.9.9)$$

$$\beta_{T_5}(\delta)=1-\Phi\left[\frac{\delta a_0+(N+1)^{-\frac{1}{2}}z_\alpha\left(\frac{\pi^2}{3}-1\right)-H(\delta)}{(N+1)^{-\frac{1}{2}}(1-\delta)\left(\frac{\pi^2}{3}-1\right)^{\frac{1}{2}}}\right]; \quad (9.9.10)$$

где $a_0=1-C-\ln(N+1)$, $H(\delta)=\delta\ln\delta+(1-\delta)\ln(1-\delta)$. Легко видеть, что для конечности мощностей при $\delta\to 0$ требуются разные скорости сходимости альтернатив для разных тестов: для T_1^2 $\delta(N)=0(N^{-\frac{3}{4}})$, для T_5 $\delta(N)=0([\sqrt{N}\ln N]^{-1})$. Это, в частности, означает, что при конечных N следует ожидать

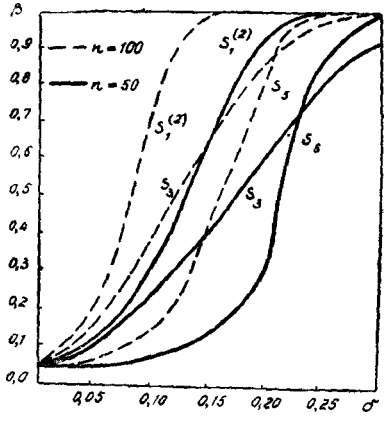

Рис. 9.9.1

предпочтительности T_1^2 перед T_5. Графики $\beta_T(\delta)$, приведенные на рис. 9.9.1, вычисленные при $N=50$, подтверждают это.

§ 9.10. ЭКСПЕРИМЕНТАЛЬНОЕ СРАВНЕНИЕ МОЩНОСТЕЙ НЕКОТОРЫХ ТЕСТОВ НА ВЫБОРОЧНЫХ ИНТЕРВАЛАХ

Трудности, связанные с теоретическим исследованием мощностных свойств тестов структуры D (см. § 9.9), можно обойти, прибегнув к экспериментальному сравнению этих тестов. Методика экспериментирования остается той же, что и описанная в § 9.8; специфика заключается лишь в вычислении тестовых статистик.

В качестве примера приведем результаты моделирования некоторых тестов для альтернатив сдвига и масштаба при разных типах распределений. Сравнению подвергались следующие тесты, упомянутые в § 9.4:

T_1^2 — тест Гринвуда, $T_1^2=\Sigma\Delta_i^2$;

T_2 — тест Кендалла-Шермана, $T_2=\sum\left|\Delta_i-\dfrac{1}{N+1}\right|$;

T_3 — тест Дарлинга, $T_3=\Sigma\ln\Delta$;

T_5 — тест Кейла-Геберта, $T_5=\Sigma\Delta_i\ln\Delta_i$.

Указанные статистики обладают асимптотической нормальностью при гипотезе; поэтому их критические значения $T(\alpha)$ при заданном объеме выборки N приближенно могут быть записаны в виде следующих выражений (z_α — квантиль уровня α стандартного нормального распределения):

$$T_1^2(\alpha)=\frac{2}{N+1}\left[\frac{z_\alpha}{\sqrt{N+1}}+1-\frac{1}{N+1}\right],$$

$$T_2(\alpha)=\frac{1}{e}\left[1+z_\alpha\sqrt{\frac{2e-5}{N}}\right],$$

$$T_3(\alpha)=-(N+1)(\ln N+C)+z_\alpha\sqrt{(N+1)\left(\frac{\pi^2}{6}-1\right)},$$

$$T_5(\alpha)=(1+C)-\ln(N+1)+z_\alpha\sqrt{\frac{\frac{\pi^2}{3}-1}{N+1}}.$$

Рассмотрим сначала альтернативу сдвига, $G(x)=F(x-a)$. Выборочные интервалы при истинности альтернативы определяются как разности между соседними порядковыми статистиками, $\Delta_i=y_{(i)}-y_{(i-1)}$; $i=1, 2, \ldots, N+1$, $y_{(0)}=0$, $y_{(N+1)}=1$; $y_{(i)}=F[F^{-1}(u_{(i)})+a]$; $u_{(i)}$ — порядковые статистики из равномерного в [0, 1] распределения (датчики равномерно распределенных чисел имеются на каждой ЭВМ). Например, для альтернативы сдвига при распределении Лапласа («двойном экспоненциальном»)

$$y_i=\begin{cases} u_i e^a, & 0\leqslant u_i<\frac{1}{2}e^{-a},\\ 1-\frac{e^{-a}}{4u_i}, & \frac{1}{2}e^{-a}\leqslant u_i<\frac{1}{2},\\ 1-e^a(1-u_i), & \frac{1}{2}\leqslant u_i\leqslant 1. \end{cases}$$

Приведем (табл. 9.10.1—9.10.3) результаты для $N=50$ и $\alpha=0{,}05$ (эксперименты показывают, что основные выводы, которые будут приведены ниже, сохраняются и при других объемах выборки и уровнях значимости).

Таблица 9.10.1

Логистическая альтернатива сдвига

T_i \ a	0,0	0,2	0,4	0,6	0,7	0,9	1,0	1,1	1,2	1,4	1,6
T_1^2	0,03	0,04	0,08	0,16	0,32	0,44	0,53	0,68	0,78	0,82	0,90
T_2	0,03	0,08	0,06	0,10	0,18	0,32	0,42	0,59	0,67	0,85	0,89
T_3	0,05	0,09	0,08	0,13	0,22	0,24	0,36	0,47	0,61	0,76	0,85
T_5	0,02	0,10	0,08	0,22	0,33	0,52	0,60	0,71	0,81	0,92	0,98

Для масштабной альтернативы ($G(x)=F[(1+\theta)x]$) выборочные интервалы определяются как разность между по-

Таблица 9.10.2

Лапласовская альтернатива сдвига

T_i \ a	0,0	0,4	0,6	0,8	0,9	1,0
T_1^2	0,03	0,23	0,51	0,76	0,87	0,99
T_2	0,03	0,20	0,45	0,71	0,84	0,98
T_3	0,05	0,17	0,40	0,62	0,75	0,96
T_5	0,02	0,27	0,57	0,82	0,89	1,00

Таблица 9.10.3

Альтернатива сдвига Коши

T_i \ a	0,0	0,4	0,6	0,8	0,9	1,1	1,2	1,4
T_1^2	0,03	0,16	0,38	0,51	0,68	0,82	0,86	0,94
T_2	0,03	0,13	0,23	0,38	0,54	0,75	0,82	0,91
T_3	0,05	0,12	0,23	0,30	0,40	0,66	0,75	0,88
T_5	0,02	0,18	0,40	0,52	0,67	0,85	0,90	0,97

рядковыми статистиками $y_{(i)}=F\left[\frac{1}{1+\theta}F^{-1}(u_{(i)})\right]$. Например, для распределения Лапласа

$$y(i)=\begin{cases}\frac{1}{2}(2u_i)^{\frac{1}{1+\theta}}, & 0\leqslant u_i\leqslant\frac{1}{2},\\ 1-\frac{1}{2}[2(1-u_i)]^{\frac{1}{1+\theta}}, & \frac{1}{2}<u_i\leqslant 1.\end{cases}$$

Приведем результаты эксперимента для $N=50$ и $\alpha=0{,}01$ (табл. 9.10.4—9.10.5).

Таблица 9.10.4

Логистическая масштабная альтернатива

T_i \ θ	0,0	0,2	0,4	0,6	0,8	0,9	1,0	1,1	1,2	1,3	1,4
T_1^2	0,02	0,04	0,07	0,15	0,42	0,61	0,67	0,83	0,92	0,96	0,99
T_2	0,01	0,02	0,04	0,06	0,16	0,24	0,32	0,45	0,60	0,75	0,81
T_3	0,01	0,02	0,03	0,05	0,15	0,19	0,24	0,33	0,46	0,61	0,66
T_5	0,02	0,03	0,05	0,10	0,32	0,50	0,58	0,79	0,87	0,95	0,96

Таблица 9.10.5

Лапласовская масштабная альтернатива

T_i \ θ	0,0	0,4	0,6	0,8	0,9	1,0	1,1	1,2	1,3	1,4	1,6
T_1^2	0,02	0,07	0,08	0,23	0,29	0,35	0,51	0,69	0,73	0,78	0,92
T_2	0,01	0,03	0,06	0,10	0,12	0,18	0,23	0,32	0,39	0,49	0,71
T_3	0,01	0,02	0,08	0,08	0,10	0,15	0,18	0,26	0,34	0,45	0,62
T_5	0,02	0,05	0,10	0,15	0,21	0,28	0,40	0,58	0,66	0,69	0,88

3. Нормальная масштабная альтернатива.

Для разнообразия и наглядности приведем экспериментальные данные для этого случая не в табличной, а в графической форме (см. рис. 9.10.1).

Несмотря на неизбежный статистический разброс экспериментальных данных, на их основе можно сделать вполне определенные выводы.

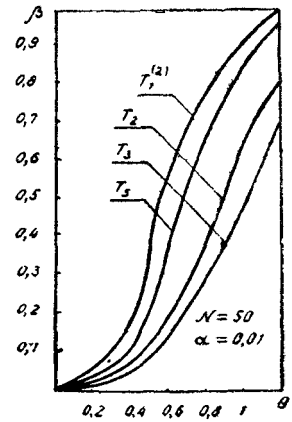

Рис. 9.10.1

1. Рассмотренные тесты имеют следующий порядок предпочтительности по мощностным свойствам (используем знак «$\succ$» в смысле «лучше»):

для симметричных* альтернатив сдвига $T_5 \succ T_1^2 \succ T_2 \succ T_3$;

для масштабных альтернатив $T_1^2 \succ T_5 \succ T_2 \succ T_3$.

2. Для масштабных альтернатив мощность рассмотренных тестов падает по мере перехода к распределениям с более затянутыми хвостами.

3. Как и критерии структуры d, критерии структуры D часто проявляют устойчивость мощности при смене распределений с сохранением энтропии канонической альтернативы. Например, для $H = -0{,}20$ получены следующие экспериментальные результаты при масштабной альтернативе (табл. 9.10.6).

Эта же особенность проявляется и для таких резко отличающихся распределений, как экспоненциальное и лапласовское: при масштабной альтернативе приравнивание энтропий этих

* Не приведенные здесь предварительные данные показывают, что для несимметричных распределений порядок предпочтительности может измениться

Таблица 9.10.6

H	Распределение	θ	Статистики			
			T_1^2	T_2	T_3	T_4
	Нормальное	0,7	0,70	0,32	0,22	0,55
—0,2	Лапласа	1,2	0,69	0,32	0,25	0,54

распределений в канонической форме обеспечивает практическое совпадение мощностей.

4. Суждения о порядке предпочтительности тестов по их мощности являются условными в том смысле, что этот порядок может изменяться при изменении класса рассматриваемых альтернатив. Однако, если практик не желает слишком большой детализации, то он может руководствоваться тем, что почти всегда T_1^2-тест Гринвуда оказывается в группе тестов с высокими мощностями.

§ 9.11. О НЕКОТОРЫХ СПОСОБАХ РЕШЕНИЯ ЗАДАЧИ СОГЛАСИЯ С ИСПОЛЬЗОВАНИЕМ СВОЙСТВ ПОРЯДКОВЫХ СТАТИСТИК

До сих пор речь шла о проверке гипотезы согласия с помощью сравнения гипотетического распределения $F(x)$ с истинным распределением $G(x)$. Считая $G(x)$ неизвестным, мы прибегаем к его непараметрическому оцениванию, а затем тем или иным способом устанавливаем приемлемость или неприемлемость расхождения между $F(x)$ и $G_N(x)$ в предположении истинности гипотезы.

Однако гипотезу согласия можно сформулировать не только в виде $F(x)=G(x)$, но и через равенства квантильных функций: $F^{-1}(u)=G^{-1}(u)$, $u\in[0, 1]$. Если теперь удастся осуществлять непараметрическое оценивание функции $G^{-1}(u)$, то возникнет возможность построения целого класса критериев согласия, обладающих своей спецификой. Такую возможность представляют свойства порядковых статистик.

Пусть $x_1, \ldots, x_N$ — выборка независимых и одинаково распределенных случайных величин с общим непрерывным и строго монотонным распределением $G(x)$; $x_{(1)}, \ldots, x_{(N)}$ — упорядоченная статистика (вариационный ряд). Легко показать, что для любого R (см. гл. III)

$$EG(x_{(R)})=\frac{R}{N+1}. \tag{9.11.1}$$

Действительно, так как $y_i=G(x_i)$, $i=1, \ldots, N$ есть выборка

из равномерного в [0, 1] распределения, то плотность распределения $f(y_{(R)})$ R-й порядковой статистики $y_{(R)}$ равна

$$f(y_{(R)}) = \frac{N!}{(R-1)!\,(N-R)!}\, y_{(R)}^{R-1}(1-y_{(R)})^{N-R}$$

и

$$EG(x_{(R)}) = Ey_{(R)} = \int_0^1 y_{(R)}\, f(y_{(R)})\, dy_{(R)} = \frac{R}{N+1}.$$

С другой стороны, дисперсия величины $G(x_{(R)})$ равна

$$DG(x_{(R)}) = Ey_{(R)}^2 - E^2 y_{(R)} = \frac{R(N-R+1)}{(N+1)^2(N+2)}. \quad (9.11.2)$$

Отсюда следует сходимость $G(x_{(R)})$ к $R/(N+1)$ в среднеквадратичном:

$$E\left(G(x_{(R)}) - \frac{R}{N+1}\right)^2 = \frac{1}{N}\sum_{R=1}^{N} \frac{R(N-R+1)}{(N+1)^2(N+2)} < \frac{1}{N}.$$

Это позволяет утверждать, что с ростом N приближенное равенство

$$G(x_{(R)}) = \frac{R}{N+1} \quad (9.11.3)$$

будет выполняться со все большей точностью. Если теперь разрешить (9.11.3) относительно $x_{(R)}$, то получится оценка неизвестной квантильной функции

$$x_{(R)} = G^{-1}\left(\frac{R}{N+1}\right) \doteq G^{-1}\left(\frac{R}{N+1}\right), \quad (9.11.4)$$

которая, в силу предположенной однозначности $G^{-1}(u)$ и (9.11.3), является состоятельной.

Имеются по крайней мере две возможности использования оценки (9.11.4) для установления сходства или различия между $F^{-1}(u)$ и $G^{-1}(u)$. Рассмотрим сначала одну из них, которая основана на геометрических свойствах совокупности пар чисел $[\varphi(x_{(R_1)}, \ldots, x_{(R_k)})]$, $\varphi[F^{-1}\left(\frac{R_1}{N+1}\right), \ldots, F^{-1}\left(\frac{R_k}{N+1}\right)]$, где $\varphi(\xi_1, \ldots, \xi_k)$ — некоторая вещественная рациональная функция, а $(R_1, \ldots, R_k)$ — набор k различных целых чисел, $R_j \in [1, N]$, $j=1, \ldots, N$. (Везде в дальнейшем будем предполагать однозначность функций $F^{-1}(u)$ и $G^{-1}(u)$).

Лемма. При $N\to\infty$, $\lim\limits_{N\to\infty} R_j/(N+1) = u_j \in (0, 1)$,

$$\lim_{N\to\infty} \varphi(x_{(R_1)}, \ldots, x_{(R_k)}) = \varphi\left(G^{-1}\left(\frac{R_1}{N+1}\right), \ldots, G^{-1}\left(\frac{R_k}{N+1}\right)\right) \quad (9.11.5)$$

Доказательство следует из теоремы сходимости рациональных функций от сходящихся последовательностей (Уилкс [1], теорема 4.3.5; Крамер [1], § 20.6).

Если теперь ввести ортогональную систему координат VOW, по оси абсцисс которой отложить величину $v = \varphi(x_{(R_1)}, \ldots, x_{(R_k)})$, а по оси ординат — величину $w = \varphi\left[F^{-1}\left(\frac{R_1}{N+1}\right), \ldots, F^{-1}\left(\frac{R_k}{N+1}\right)\right]$, то в силу леммы двумерная точка (v, w) будет при $N \to \infty$ сходиться в вероятностном смысле к точке $[\varphi(G^{-1}(u_1), \ldots, G^{-1}(u_k)), \varphi(F^{-1}(u_1), \ldots, F^{-1}(u_k))]$, которая лежит на некоторой параметрической кривой в плоскости VOW. В этом смысле мы будем для краткости говорить, что двумерная точка (v, w) сходится к некоторой кривой (Тарасенко, Шуленин [4]).

Рассмотрим теперь несколько примеров, отличающихся заданием функции φ.

Пример 9.11.1. Если $\varphi(\xi) = \xi$, то кривая, к которой сходятся точки $\left(v_R = x_{(R)}, w_R = F^{-1}\left(\frac{R}{N+1}\right)\right)$, есть $w = F^{-1}G(v)$. Действительно, параметрическое задание кривой в виде $[v = G^{-1}(u), w = F^{-1}(u)]$, к которой в указанном выше смысле сходятся эмпирические точки (v_R, w_R), выражает именно эту функцию. Из этого простого факта вытекает ряд интересных следствий.

Следствие 1. Условие $F = G$ является необходимым и достаточным для сходимости точек (v_R, w_R), $R = \overline{1, N}$ к прямой $w = v$.

Следствие 2. Если F и G принадлежат к одному типу, т. е. $G(x) = F(\lambda x + \mu)$, то точки (v_R, w_R) сходятся к прямой $w = \lambda v + \mu$. В самом деле, $F^{-1}G(v) = F^{-1}F(\lambda v + \mu) = \lambda v + \mu$.

Два последних утверждения хорошо известны и широко используются на практике для проверки соответствия истинного и предполагаемого распределений с помощью «вероятностной бумаги», на которой заранее нанесена разметка по одной из осей, определяемая функцией $F^{-1}(u)$. Правда, обычно на вероятностную бумагу наносится эмпирическая функция распределения ($F_N = R/N$), но при $N \gg 1$ разницей между R/N и $R/(N+1)$ можно пренебречь.

Для семейства Ω непрерывных распределений с симметричными плотностями можно получить дальнейшие результаты. Напомним (§ Д.4), что если $F \in \Omega$ имеет более затянутые хвосты, чем $G \in \Omega$, то $F^{-1}(u) = a(u) G^{-1}(u)$, где $a(u)$ — неубывающая функция $u \in (1/2, 1)$. Отметим, что всегда существует такая точка $u_0 \in (1/2, 1)$, что при всех $u > u_0$ $a(u) > 1$. Теперь может быть доказано

Следствие 3. Если $F \in \Omega$ имеет более (менее) затяну-

тые хвосты, чем $G \in \Omega$, то, начиная с некоторого $v_0 = v(u_0)$, кривая $w = F^{-1}G(v)$ будет вогнутой (выпуклой) неубывающей функцией, лежащей ниже (выше) прямой $w = v$.

Доказательство. Пусть существуют f и g — плотности для F и G соответственно. Тогда

$$\frac{d}{dx} F^{-1} G(x) = \frac{g(x)}{f[F^{-1}G(x)]} = \frac{g[G^{-1}(u)]}{f[a(u)\, G^{-1}(u)]}. \quad (9.11.6)$$

В силу свойств функции $a(u)$ видно, что при $u > u_0$ эта производная всегда больше единицы, что и доказывает основное утверждение следствия 3. Утверждение в скобках доказывается аналогично.

Для решения задачи согласия со сложной гипотезой $H_0 : F\left(\frac{x-\theta}{\sigma}\right)$ желательно установление фактов, которые бы не были связаны с параметрами θ и/или σ гипотезы. Такими свойствами, очевидно, будут обладать разности и отношения порядковых статистик и квантилей. Это приводит нас к другим примерам функций φ.

Пример 9.11.2. Если $\varphi(\xi_1, \xi_2) = \xi_1 - \xi_2$, т. е. $v_{RK} = x_{(R)} x_{(K)}$, $w_{RK} = F^{-1}\left(\frac{R}{N+1}\right) - F^{-1}\left(\frac{K}{N+1}\right)$, $1 < K \leqslant R < N$, то при $N \to \infty$, $R/(N+1) \to u_R$, $K/(N+1) \to u_k$, $0 < u_k < u_R < 1$,

$$v_{RK} \to G^{-1}(u_R) - G^{-1}(u_k), \quad w_{RK} \to F^{-1}(u_R) - F^{-1}(u_k). \quad (9.11.7)$$

Для симметричных распределений можно продвинуться дальше и утверждать, что

$$v_{RK} \to a(u_R)\left[w_{Rk} + \left(1 - \frac{a(u_k)}{a(u_R)}\right) F^{-1}(u_k)\right]. \quad (9.11.8)$$

Это соотношение вытекает из (9.11.7) с учетом того, что $G^{-1}(u) = a(u) F^{-1}(u)$.

Следствие 1. Если F и G принадлежат одному типу, то есть отличаются только сдвигом и/или масштабом, $G(x) = F^{-1}(\lambda x + \mu)$, то точки (v_{Rk}, w_{Rk}) сходятся к прямой $w = u/\lambda$, так как при этом $a(u) = \lambda = \text{const}$.

Следствие 2. Если G имеет более (менее) затянутые хвосты, чем F, то, начиная с некоторого u_0, точки (v_{Rk}, w_{Rk}) будут располагаться выше (ниже) соответствующей прямой.

Пример 9.11.3. Пусть $\varphi(\xi_1, \xi_2, \xi_3, \xi_4) = (\xi_1 - \xi_2)/(\xi_3 - \xi_4)$. Тогда

$$v_{RKLM} = (x_{(R)} - x_{(K)})/(x_{(L)} - x_{(M)}), \quad w_{RKLM} =$$

$$= \left(F^{-1}\left(\frac{R}{N+1}\right) - F^{-1}\left(\frac{K}{N+1}\right)\right) \Big/ \left(F^{-1}\left(\frac{L}{N+1}\right) - F^{-1}\left(\frac{M}{N+1}\right)\right),$$

$$R > K, \quad L > M.$$

Если сохраняются предположения о распределениях F и G и об асимптотическом поведении последовательностей номеров R, K, L, M, $0<u_K<u_R<1$, $0<u_M<u_L<1$, то точки (v_{KLM}, w_{RKLM}) сходятся к прямой $v=w$, если и только если F и G принадлежат к одному типу, т. е. $G(x)=F\left(\frac{x-\theta}{\sigma}\right)$ для всех x и любых θ и $\sigma>0$.

Следствие. Если F и G симметричны и принадлежат разным типам, то для конкретных отношений между R, K, L, M (u_R, u_K, u_L, u_M) можно установить, с какой стороны от прямой $v=w$ будут (в среднем) располагаться точки (v_{RKLM}, w_{RKLM}). Например, если $R>L>M>K$ и G имеет более затянутые хвосты, чем F, то w_{RKLM} статистически больше v_{RKLM}.

Перейдем теперь к вопросам использования перечисленных выше свойств порядковых статистик и их комбинаций. Отметим основные особенности этого подхода к статистическим задачам.

Во-первых, сам собой напрашивается класс процедур для проверки гипотезы о согласии, основанных на оценке близости соответствующих эмпирических точек к прямой. Характерно, что количественное определение «прямизны» оказывается хотя и возможным, но при практическом использовании громоздким (например, по методу наименьших квадратов). Часто для практики оказывается полезной аппелляция к геометрическому восприятию человека: одного взгляда на последовательность точек бывает достаточно, чтобы сказать, приемлема ли ее аппроксимация прямой линией. Такие способы, несмотря на их «неформальность» и нестрогость, широко употребляются в практике и даже имеют свое название — «быстрые процедуры». К сожалению, при этом остаются неопределенными вероятности ошибок.

Во-вторых, ценной особенностью представления экспериментальных данных с помощью методик, описанных выше, является то, что проверка гипотезы согласия с их помощью позволяет не только принять или отвергнуть гипотезу, но и при ее отвержении дает информацию об относительной затянутости хвостов неизвестной альтернативы по сравнению с гипотезой. Это существенно может облегчить определение истинного распределения (что часто и является целью использования критериев согласия) за счет целенаправленного перебора гипотез.

Для иллюстрации этого свойства на рис. 9.11.1 приведены результаты эксперимента с пятью типовыми распределениями. По оси абсцисс отложена величина v_{RKLM} из примера 9.11.3, по оси ординат — w_{RKLM}. В качестве F взято логисти-

ческое распределение. Объем выборки 200 (на графике нанесены не все точки), номера кривых соответствуют: *1* — равномерному распределению, *2* — нормальному, *3* — логистическому, *4* — Лапласа, *5* — Коши. $R=N+1-i$, $K=i$, $L=\frac{N}{2}+i$, $M=\frac{N}{2}+1-i$, $i=1, 2,\ldots, \frac{N}{2}$.

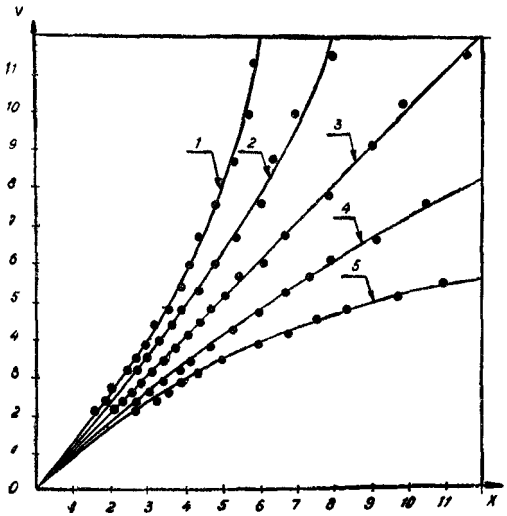

Рис 9.11 1

Наконец отметим, что использование результатов примеров 9.11.1—9.11.3 позволяет, при необходимости, как исключить параметры сдвига и/или масштаба гипотетического распределения, так и, наоборот, оценивать эти параметры. Из этих соображений можно предложить конкретные статистики для оценки параметров сдвига или масштаба при истинности гипотезы. Например, несмещенной оценкой масштаба при произвольном сдвиге может служить статистика

$$\hat{\sigma}_N=\frac{1}{N(N+1)}\sum_{R\neq K}\frac{x_{(R)}-x_{(K)}}{F^{-1}\left(\frac{R}{N+1}\right)-F^{-1}\left(\frac{K}{N+1}\right)},$$

а для сдвига при произвольном масштабе имеем:

$$\hat{\theta}_N=\frac{1}{N(N+1)}\sum_{R\neq K}\frac{F^{-1}\left(\frac{R}{N+1}\right)x_{(R)}-F^{-1}\left(\frac{K}{N+1}\right)x_{(K)}}{F^{-1}\left(\frac{R}{N+1}\right)-F^{-1}\left(\frac{K}{N+1}\right)}.$$

Как уже отмечалось выше, для решения задачи согласия можно предложить не только «быстрые», неформализованные методы, описанные выше, но и построить ряд критериев алгоритмически. Тесты, использующие свойства порядковых статистик, будут возникать, если вводить «расстояния» между гипотезой и альтернативой через их обратные функции, F^{-1} и G^{-1}, а оценки этих расстояний строить путем оценивания G^{-1} через порядковые статистики. Как и при построении тестов других типов, «расстояния» $\rho(F^{-1}, G^{-1})$ могут быть введены различным образом, так что возникает целый класс тестов. Для примера приведем некоторые «расстояния» ρ и порождаемые ими тестовые статистики S.

ρ	S
$\int_0^1 [G^{-1}(u)-F^{-1}(u)]^r du$	$\sum_R \left[x_{(R)} - F^{-1}\left(\frac{R}{N+1}\right)\right]^r$
$\int_0^1 \frac{G^{-1}(u)-G^{-1}(1-u)}{F^{-1}(u)-F^{-1}(1-u)} du$	$\sum_R \frac{x_{(R)}-x_{(N+1-R)}}{F^{-1}\left(\frac{R}{N+1}\right)-F^{-1}\left(1-\frac{R}{N+1}\right)}$
$\int_{1/2}^1 [G^{-1}(u)/F^{-1}(u)]\,du$	$\sum_{R=[(N+1)/2]} x_{(R)}/F^{-1}(R/(N+1))$
$\int_{1/2}^1 \ln [G^{-1}(u)/F^{-1}(u)]\,du$	$\sum_{R=\left[\frac{N+1}{2}\right]}^{N} \ln \left[x_{(R)}\, F^{-1}\left(\frac{R}{N+1}\right)\right]$
$\sup_u [G^{-1}(u)-F^{-1}(u)]$	$\sup_R \left[x_{(R)}-F^{-1}\left(\frac{R}{N+1}\right)\right]$.

Свойства тестов данного класса пока не изучены. Некоторые соображения по поводу статистики, основанной на «расстоянии» $\rho = \int_{1/2}^1 \frac{G^{-1}(u)-G^{-1}(1-u)}{F^{-1}(1-u)}\,du$, имеются у Гаека [4].

§ 9.12. О КРИТЕРИЯХ СОГЛАСИЯ ДЛЯ СЛОЖНОЙ ГИПОТЕЗЫ

Пожалуй, наиболее интересной для практики является задача согласия со сложной гипотезой, когда делается предположение лишь о принадлежности истинного распределения к определенному классу распределений с произвольными параметрами: $H_0: F(x) = F(x/\theta_1, \ldots, \theta_k)$.

К настоящему моменту развито несколько различных подходов к решению этой задачи. Исторически первым является подход, при котором неизвестные параметры $\{\theta_i\}$ гипотетического распределения оцениваются по самой выборке, а затем решается задача согласия с простой гипотезой $H_0: F(x) = F(x/\hat{\theta}_1, \ldots, \hat{\theta}_k)$. Особенность при этом, оказывается, состоит в том, что не всякие оценки параметров обеспечивают необходимые качества такого теста. Второй подход основывается на том, что некоторые типы распределений допускают построение статистик, статистически нечувствительных к изменениям параметров, но обнаруживающих переход к распределениям другого типа. В отличие от первого подхода, здесь речь идет пока лишь о типах, определяемых произвольными сдвигом и масштабом. Так, классы распределений Рэлея-Райса и γ-распределений могут служить нулевой гипотезой при первом подходе, но не при втором. Третий подход основан на том, что можно найти оценку $F_N(x/\hat{\theta}_1, \ldots, \hat{\theta}_k)$, которая в ряде случаев оказывается лучшей, чем используемая при первом подходе оценка $F(x/\hat{\theta}_1, \ldots, \hat{\theta}_k)$. Четвертый подход состоит в том, чтобы, используя некоторые свойства выборочных моментов и остроумную рандомизацию, спроектировать сложную гипотезу на простую.

Кратко изложим результаты, полученные в указанных четырех направлениях.

Основная идея первого подхода состоит в том, чтобы рассматривать вместо «расстояния» $\rho(F, G)$ меру уклонения $\hat{\rho}(F(x/\hat{\theta}_1, \ldots, \hat{\theta}_k), G)$; тестовые статистики при этом будут получаться путем оценивания $\hat{\rho}$ при использовании вместо G эмпирической функции распределения G_N. Практическое значение получаемых таким образом тестов зависит от ответов на следующие вопросы:

1. При каких условиях тесты на статистиках $\hat{\rho}_N$ обладают непараметричностью? Точнее, какие требования должны быть предъявлены к оценкам $\{\theta_i\}$, чтобы асимптотические распределения при гипотезе $F(x/\hat{\theta}_1, \ldots, \hat{\theta}_k)$ совпадали с таковыми при гипотезе $F(x/\theta_1, \ldots, \theta_k)$? (Это гарантировало бы сохранение уровня значимости и позволило бы пользоваться уже имеющимися таблицами критических значений).

2. Если таких условий не существует или они выполняются в очень частных случаях, то каковы предельные распределения статистик ρ_N при заданном типе оценок $\{\hat{\theta}_i\}$?

При каких ограничениях эти предельные распределения не зависят от конкретных значений параметров $\{\theta_i\}$? (Тогда появилась бы возможность составления таблиц специально для этих тестов).

Поиску ответов на эти вопросы посвящены работы Дарлинга [3], Гихмана, Гнеденко и Смирнова [1], Лиллифорса [1, 2]. Ими были рассмотрены модификации тестов Колмогорова (D_N) и Крамера — Мизеса — Смирнова (ω^2). Оказалось, что для модифицированного ω^2-теста предельное распределение совпадает с таковым для простой гипотезы, только если параметры $\{\theta_i\}$ допускают «суперэффективные» оценки, то есть если $NE[(\hat{\theta}_i - \theta_i)^2] \to 0$ при $N \to \infty$. Существование суперэффективных оценок является чрезвычайным событием; на практике обычно встречаются оценки, для которых $\sqrt{N}(\hat{\theta}_i - \theta_i)$ имеет предельное нормальное $(0, \sigma^2)$ распределение. Для таких оценок модифицированные ω^2- и D_N-тесты теряют свойство непараметричности. Предельные распределения при гипотезе в этом случае совпадают с распределением величины $\int Y^2(t)\,dt$, где $Y(t)$ — определенный гауссов процесс и зависит от типа распределения при нулевой гипотезе и от истинных значений параметров. Только в том случае, когда параметры являются сдвигом и масштабом и их оценки асимптотически эффективны (например, оценки максимального правдоподобия), исчезает зависимость предельного распределения статистики от этих параметров. В связи с этим Кац, Кифер и Вольфовиц рассмотрели критерий нормальности: параметры нормального распределения оказываются именно параметрами сдвига и масштаба (немаловажна также широкая распространенность нормальных распределений во многих практических случаях). Они довели результаты до получения таблиц предельных распределений для модифицированных ω^2- и D-тестов.

Несколько слов об использовании χ^2-теста. Уже сам Пирсон [1] сделал попытку обобщить χ^2-тест на случай сложной гипотезы путем замены истинных математических ожиданий m_i на их оценки, полученные с помощью предварительного оценивания неизвестных параметров $\{\theta_i\}$ по самой выборке. Ему, однако, не удалось полностью решить вопрос о правильном выборе числа степеней свободы в этом случае. Этот вопрос был решен Фишером [1]. Последний подчеркнул, что предельное распределение χ^2-статистики существенно определяется тем, какой метод выбран для оценивания параметров. Если берутся оценки максимального правдоподобия, то верна следующая теорема (Крамер [1]):

Теорема 9.12.1. Пусть вероятности $p_i(\theta_1, \ldots, \theta_s)$,

$i=1, \ldots, k$, известным образом зависят от $s<k$ параметров $\theta_1, \ldots, \theta_s$. Пусть для всех точек невырожденного интервала A в s-мерном пространстве $(\theta_1, \ldots, \theta_s)$ вероятности $\{p_i\}$ удовлетворяют следующим требованиям:

а) $\sum_{i=1}^{k} p_i = 1$;

б) $p_i > C^2 > 0$ для всех i;

в) каждое из $\{p_i\}$ имеет непрерывные производные

$$\frac{\partial p_i}{\partial \theta_j} \text{ и } \frac{\partial^2 p_i}{\partial \theta_j \partial \theta_h};$$

г) матрица $D = \left\{\frac{\partial p_i}{\partial \theta_j}\right\}$ имеет ранг s.

Тогда X^2 в пределе, при $N \to \infty$, имеет χ^2-распределение с $(k-s-1)$ степенями свободы.

Таким образом, χ^2-критерий оказывается асимптотически непараметрическим, если пользоваться оценками максимального правдоподобия для неизвестных параметров (при этом лишь число степеней свободы уменьшается на число оцениваемых параметров). Это, естественно, делает χ^2-критерий весьма привлекательным для практики именно в случае сложной гипотезы. Правда, оценки максимального правдоподобия не всегда оказываются существующими или достаточно просто вычислимыми, но это особый вопрос.

Принципиально иной подход к решению задачи согласия со сложной гипотезой состоит в том, чтобы сконструировать «расстояние» (или сразу статистику), заведомо не зависящее от неизвестных параметров, но изменяющееся при переходе от распределений одного типа к другому. Иными словами, вводимая мера уклонения между гипотезой и альтернативой должна быть функционалом этих распределений, инвариантным относительно неизвестных параметров. К настоящему времени принципы построения таких «расстояний» (или непосредственно статистик) разработаны лишь по отношению к параметрам сдвига и масштаба. Эти принципы состоят в использовании свойств выборочных интервалов и леммы Вайсса (см. § D.4) для исключения сдвига и в использовании операции деления для исключения масштабного параметра.

Например, Вайсс [4] предложил пользоваться статистиками вида

$$W_N(r) = \sum \left\{ N\Delta_R f \left| F^{-1}\left(\frac{R}{N}\right)\right|\right\}^r, \qquad (9.12.1)$$

в которых сдвиг уже исключен, а тестовую статистику Вайсс конструирует в виде

$$Z_N = N \frac{W_N(2)}{[W_N(1)]^2}. \qquad (9.12.2)$$

Вайсс доказал состоятельность Z_N-теста и привел соображения по поводу асимптотической нормальности статистики (9.12.2) при нулевой гипотезе. Блюменталь [3], в свою очередь, предложил пользоваться тестовыми статистиками вида:

$$P_N(r) = W_N(r)\, W_N(-r), \qquad (9.12.3)$$

которые также инвариантны к сдвигу и масштабу. Исследуя асимптотические распределения статистик $P_N(r)$ при гипотезе, он обнаружил, что эти распределения не зависят от F лишь при $\frac{1}{2} \leqslant r \leqslant 1$. При $0 < r < \frac{1}{2}$ распределения являются асимптотически нормальными с параметром σ_1^2, сложным образом зависящим от F. Блюменталь показал также, что при проверке на экспоненциальность и Вейбулловских альтернативах $\mathrm{ARE}\left[P_N\left(\frac{1}{2}\right), P_N\left(r \in \left(0, \frac{1}{2}\right)\right)\right] = 0$, что предельная мощность $P_N\left(r \in \left(0, \frac{1}{2}\right)\right)$-тестов совпадает с таковой для колмогоровского теста и что χ^2-тест имеет нулевую эффективность по сравнению с $P_N(r)$-тестом. Однако ему не удалось довести (при $1/2 \leqslant r \leqslant 1$) вычисление асимптотических распределений до вида, допускающего табулирование, так что практическое использование этого класса тестов пока невозможно. К этому же типу процедур можно отнести тесты, основанные на отношениях выборочных интервалов, подобные тем, которые введены в конце § 9.11.

Третий подход к задаче согласия со сложной гипотезой предложен Сринивасаном [1]. Отличие от первого подхода состоит в том, что вместо оценки $F(x/\hat{\theta}_1, \ldots, \hat{\theta}_k)$ использовать другую оценку, $\tilde{F}(x/\theta_1, \ldots, \theta_k)$.

Очевидно, что $z = c(u - x_1)$, где $c(t) = \{1 : t \geqslant 0;\ 0 : t < 0\}$ — функция сравнения, является несмещенной оценкой для $F(x/\theta_1, \ldots, \theta_k)$, если x_1 — выборочное значение из этого распределения: $Ez = F(x/\theta_1, \ldots, \theta)$. Пусть $t_1, \ldots, t_k$ являются достаточными статистиками для $\theta_1, \ldots, \theta$ соответственно. Тогда из общей теории достаточных статистик (см. Блэквелл [1], Леман и Шеффе [1], Рао [1]) следует, что статистика

$$\tilde{F}(u|\theta_1, \ldots, \theta_k) = \hat{E}(z|t_1, \ldots, t_k) \qquad (9.12.4)$$

является несмещенной оценкой F с меньшей дисперсией, чем у z. Теперь можно записать модифицированную статистику Колмогорова:

$$\tilde{D}_N=\sup_x |F_N(x)-\tilde{F}(x|\theta_1,\ldots,\theta_k)|. \qquad (9.12.5)$$

Сринивасан показал, что в случае нормальной и экспоненциальной гипотез с неизвестными параметрами распределение $\tilde{D}_N$ не зависит от этих параметров.

Для наглядности приведем пример с экспоненциальной гипотезой. Пусть $H_0: f(x/\lambda)=\lambda \exp(-\lambda x)$, $x\geqslant 0$, где λ — неизвестный параметр. При этом распределении выборочное среднее $t=\bar{x}=N^{-1}\Sigma x_i$ является полной достаточной статистикой для λ. Прямые вычисления дают:

$$\tilde{F}(u|\lambda)=E(z|\bar{x})=1-\left(1-\frac{u}{N\bar{x}}\right)^{N-1}.$$

Модифицированная D_N-статистика может быть записана в виде

$$\tilde{D}_N=\max_i |F_N(x_i)-\tilde{F}(x_i|\lambda)|=\max_i |F_N(y_i)-\tilde{F}(y_i|1)|,$$

где $y_i=\lambda x_i$. Этим равенством и обосновывается независимость распределения $\tilde{D}_N$ от λ. Сринивасан составил методом Монте-Карло таблицы критических значений для $N=4(1)10(2)30$ и уровней значимости $\alpha=10^{-3}$; 10^{-2}; $5\cdot 10^{-2}$; 10^{-1}; $1{,}5\cdot 10^{-1}$ и $2\cdot 10^{-1}$. Такие же таблицы приведены им для критерия нормальности при произвольных a и σ^2. Сринивасан также сделал вывод о том, что для экспоненциальных альтернатив $\tilde{D}$-тест мощнее D_N-теста, для других их мощности близки. Однако Шафер, Филькенштейн и Колинс [1] показали ошибочность этого результата: ими доказано, что и при экспоненциальных альтернативах мощности $\tilde{D}_N$- и D_N-тестов близки. Таким образом, преимущества третьего подхода пока остаются проблематичными.

Четвертый подход, в отличие от предыдущих, связан не с той или иной модификацией тестовых статистик, а с определенным преобразованием выборочных значений (Дурбин [1]). В результате такого преобразования сложная гипотеза отображается на простую. Например, если $\{x_i\}$-выборка из нормального распределения $N(a, \sigma^2)$ с неизвестными a и σ^2, $\bar{x}=N^{-1}\Sigma x_i$, $s^2=(N-1)^{-1}\Sigma(x_i-\bar{x})^2$, а $\bar{x}'$ и s'^2 — выборочные среднее и дисперсия, построенные по независимой выборке объема N из распределения $N(0, 1)$, то случайные величины $x_1', \ldots, x_N'$, определенные равенством

$$(x_i'-\bar{x}')/s'=(x_i-\bar{x})/s,$$

оказываются независимыми и распределенными нормально $N(0, 1)$, независимо от того, каковы a и σ^2. Дурбин [1] дал строгое доказательство этого и изложил общий метод полу-

чения аналогичных рандомизированных преобразований для произвольных распределений, допускающих достаточные статистики для $\theta_1, \ldots, \theta$.

§ 9.13. О ДВУВЫБОРОЧНОЙ ЗАДАЧЕ СОГЛАСИЯ. КРИТЕРИИ ОДНОРОДНОСТИ

Пусть $x_1, \ldots, x_m$ и $y_1, \ldots, y_n$ — две выборки независимых наблюдений, распределения которых ($F(x)$ и $G(y)$ соответственно) непрерывны, а в остальном — произвольны. Необходимо проверить гипотезу $H_0: F(x) = G(x)$ против альтернативы $H_1: F(x) \neq G(x)$. Такая задача получила название задачи однородности, поскольку гипотеза состоит в том, что смешанная (объединенная) выборка «однородна», т. е. взята из одного распределения. Легко видеть, что задача однородности является определенным обобщением задачи согласия на случай, когда непараметрична не только альтернатива, но и гипотеза.

Различные критерии однородности будут порождаться разными «расстояниями» $\rho(F, G)$ и разными оценками этих расстояний (так как снова могут использоваться различные оценки распределений). Легко видеть, что многообразие возможных критериев однородности даже шире, чем для критериев *согласия*: было бы интересно, например, выяснить свойства критериев однородности при использовании оценки F одного типа, а оценки G — другого. В литературе, однако, основное внимание уделено критериям однородности, являющимся обобщениями критериев Колмогорова — Смирнова и Крамера — фон Мизеса. Другие критерии представлены значительно беднее. В данном параграфе мы дадим краткий обзор некоторых основных результатов, имеющихся к данному времени.

Тесты Колмогорова-Смирнова. Для односторонних альтернатив $H_1^+: F > G$, $H_1^-: F < G$ и двусторонней альтернативы $H_1: F \neq G$ естественным является использование статистик

$$D_{mn}^{+} = \sup_x [F_m(x) - G_n(x)], \tag{9.13.1}$$

$$D_{mn}^{-} = \sup_x [G_n(x) - F_m(x)], \tag{9.13.2}$$

$$D_{mn} = \sup_x |F_m(x) - G_n(x)|, \tag{9.13.3}$$

где F_m и G_n — соответствующие эмпирические функции распределения. Смирнов [5] доказал, что асимптотические рас-

пределения величин $\sqrt{\frac{mn}{m+n}}D^{+}_{mn}$ и $\sqrt{\frac{mn}{m+n}}D_{mn}$ при $0<a\leqslant\frac{m}{n}\leqslant$ $\leqslant b<\infty$ совпадают с таковыми для $\sqrt{N}D^{+}_{N}$ и $\sqrt{N}D_{N}$. Последние подробно табулированы (Большев и Смирнов [1], Оуэн [1]). Для случая умеренных значений m и n Боровков [1] и Королюк [1] получили приближенные формулы; Большев и Смирнов [1] привели их к виду, более удобному для статистических приложений. Гнеденко с сотрудниками получили точные распределения D_{mn}, D^{+}_{mn}, D^{-}_{mn} (D_{mn} и D^{+}_{mn} при $m=n$ и $m\neq n$ — Гнеденко и Королюк [1]; совместное распределение D^{+}_{mn} и D_{mn} — Гнеденко и Рвачева [1]); Карвальхо [1] повторил результат Гнеденко и Королюка изящным методом теории случайных блужданий. Свойством статистик (9.13.1)—(9.13.3) посвящено значительное количество работ; их обзоры можно найти у Дарлинга [1], Большева и Смирнова [1], Гихмана, Гнеденко и Смирнова [1].

При больших объемах выборок вычисление статистик (9.13.1)—(9.13.3) становится громоздким делом. Возникает мысль о возможности сокращения объема вычислений за счет предварительной группировки данных, подобно тому, как это делается при построении гистограмм. Гихман [1, 2] рассмотрел вопрос об асимптотическом поведении таких модифицированных эмпирических функций распределения и статистик (9.13.1)—(9.13.3) для них. Красиво, например, выглядит результат

$$\lim_{n\to} Pr\{F_n(x)>G_n(x) \text{ для всех } x, \text{ таких, что } \alpha<U(x)<\beta\}=$$
$$=\frac{1}{\pi}\arcsin\sqrt{\frac{\alpha(1-\beta)}{\beta(1-\alpha)}},$$

где $U(x)$ — гипотетическое распределение ($=F=G$).

Значительный интерес представляет изучение мощностных свойств D_{mn}-теста. Капон [2], пользуясь результатами Мэсси [2], получил нижние границы питмановской эффективности D_{mn}-теста для ряда альтернативных распределений. Сравнивая D_{mn}-тест с L-тестами отношения правдоподобия, Капон получил следующие неравенства:

а) нормальная альтернатива сдвига:

$$0{,}637\doteq\frac{2}{\pi}\leqslant \mathrm{ARE}\,(D_{mn},\ L)\leqslant 1;$$

б) нормальная альтернатива масштаба:

$$0{,}117\doteq\frac{1}{\pi e}\leqslant \mathrm{ARE}\,(D_{mn},\ L)\leqslant 1;$$

в) распределение Коши, альтернатива сдвига:

$$0{,}811 \doteq \frac{8}{\pi^2} \leqslant \mathrm{ARE}\,(D_{mn},\ L) \leqslant 1;$$

г) распределение Коши, альтернатива масштаба:

$$0{,}203 \doteq \frac{2}{\pi^2} \leqslant \mathrm{ARE}\,(D_{mn},\ L) \leqslant 1;$$

д) лапласовская альтернатива сдвига:

$$\mathrm{ARE}\,(D_{mn},\ L) = 1;$$

е) лапласовская альтернатива масштаба:

$$0{,}541 \doteq \frac{4}{e^2} \leqslant \mathrm{ARE}\,(D_{mn},\ L) \leqslant 1.$$

При сравнении D_{mn} -теста с t-тестом Стьюдента, оптимальным для нормальной альтернативы сдвига,

$$\mathrm{ARE}\,(D_{mn},\ t_{mn}) \geqslant 4 \sup_x g^2(x), \qquad (9.13.4)$$

где $g(x)$ — плотность альтернативного распределения. Ясно, что (9.13.4) может быть сколь угодно велико. Нижняя граница (9.13.4) при условиях $\int xg\,dx = 0$ и $\int x^2 g\,dx = 1$ оказывается равной $1/3$.

Рамачандрамурти [1] проделал аналогичную работу по отношению к D_{mn} -тесту и нормальным альтернативам сдвига и масштаба и к D^{+}_{mn} -тесту и нормальной альтернативе сдвига. Он обнаружил зависимость нижней границы ARE от отношения m/n и показал, что при $m/n > 40$ $\mathrm{ARE} \geqslant 0{,}36$. Клотц [1] вычислил бахадуровскую эффективность D_{mn} - и D^{+}_{mn} -тестов для нормальных альтернатив сдвига и масштаба и привел таблицы зависимости этой эффективности от параметров альтернатив.

Тест Купера. Как и в тестах согласия, в тестах однородности критерии Колмогорова—Смирнова гораздо более чувствительны к альтернативам сдвига, нежели к масштабным альтернативам (сравни, например, нижние границы эффективности D_{mn} -теста, приведенные выше). Купер [1] предложил использовать V_N-тест согласия ($V_N = D^{+}_{N} + D^{-}_{N}$); который имеет те же, что и тесты Колмогорова — Смирнова, мощностные свойства для альтернатив сдвига, но значительно более последних чувствителен к альтернативам масштаба (см. § 9.7). Естественно ожидать, что двувыборочный аналог теста Купера будет обладать таким же свойством. Гне-

денко [3] получил асимптотическое распределение V_{mn} -статистики при истинности гипотезы ($m=n=N$, $z>0$):

$$\lim_{N\to\infty}\sqrt{2N}\,Pr\left\{V_{mn}=\sqrt{\frac{1}{2N}}\,[z\sqrt{2N}]\right\}=$$

$$=8z\sum_{j=1}^{\infty}(4j^2z^2-3j^2)\,e^{-2j^2z^2}. \qquad (9.13.5)$$

Комо [1] дает точное и асимптотическое распределение V_{mn}-статистики при гипотезе в следующей удобной форме:

$$P\left[\sqrt{\frac{\overline{N}}{2}}\,V_{mn}\geqslant D_\alpha=\frac{k}{\sqrt{2N}}\right]=$$

$$=2\sum_{i=1}^{\infty}\frac{k\binom{2N}{N+ik}-(k+1)\binom{2N}{N+i(k+1)}}{\binom{2N}{N}}, \qquad (9.13.6)$$

(здесь $m=n=N$), и

$$\lim_{m,\,n\to\infty}P\left[\sqrt{\frac{mn}{m+n}}\,V_{mn}\geqslant D_\alpha\right]=2\sum_{i=1}^{\infty}(4i^2D_\alpha^2-1)\,e^{-2i^2D_\alpha^2}. \qquad (9.13.7)$$

В этой же статье Комо высказал соображения, согласно которым питмановская эффективность V_{mn} -теста при масштабных альтернативах вдвое превышает эффективность тестов Колмогорова — Смирнова и остается той же при альтернативах сдвига.

ω^2_{mn} -тест. Естественным аналогом ω^2-теста для задачи однородностей является тест, основанный на статистике

$$W^2_{mn}=\frac{mn}{m+n}\int_{-\infty}^{\infty}[F_m(x)-G_n(x)]^2\psi\times$$

$$\times\left[\frac{mF_m(x)+nG_n(x)}{m+n}\right]d\left(\frac{mF_m(x)+nG_n(x)}{m+n}\right), \qquad (9.13.8)$$

где F_m и G_n — соответствующие эмпирические функции распределения. Розенблатт [3] показал, что при $m\to\infty$, $n\to\infty$ и $m/n\to\lambda=\text{const}>0$ предельное распределение статистики W_{mn} (при $\psi\equiv1$) совпадает с таковым для ω^2_N. Для этого распределения (при $\psi\equiv1$) среднее и дисперсия равны соответственно $1/6$ и $1/45$, тогда как

$$E\omega^2_{mn}=\frac{1}{6}\left(1+\frac{1}{m+n}\right)$$

$$D\omega_{mn}^2 = \frac{1}{45}\left(1+\frac{1}{m+n}\right)\left(1+\frac{1}{m+n}-\frac{3}{4}\,\frac{mn}{m+n}\right).$$

Поэтому при практическом использовании ω_{mn}^2 -теста рекомендуется пользоваться соответствующими поправками на конечность m и n, хотя скорость сходимости к предельному распределению весьма высока (см. Большев и Смирнов [1]). Точные таблицы для m и $n \leqslant 8$ даны Андерсоном [1]. Бурр [1] расширил эти таблицы (при $m=n=N$ до $N=10$); далее, он показал, что правый хвост ($P_N \leqslant 0,\ 1$) точного распределения P_N при $N>10$ может быть выражен через асимптотическое (P_∞) распределение эмпирической формулой

$$P_N = P_\infty\left(1+\frac{f(t)}{N}\right),$$

и дал таблицы для $f(t)$.

Некоторые общие соображения об одном методе исследования мощности ω_{mn}^2 -теста имеются у Андерсона и Дарлинга [1].

Другие тесты однородности, основанные на эмпирической функции распределения. Различие между F и G и, соответственно, между F_m и G_n может быть обнаружено не только с помощью оценки некоторого «расстояния» $\rho(F, G)$, но и за счет изменения любого свойства эмпирических ф. р. F_m и G_n при переходе от гипотезы к альтернативе. Так например, Гнеденко и Михалевич [1, 2] установили, что число пересечений F_m и G_n статистически различно при гипотезе однородности и при альтернативе. В случае истинности гипотезы число $C(m, n)$ «положительных скачков» функции $F_m(x)$, т. е. таких $\{x_i\}$, что

$$F_m(x_k-0) = \frac{k-1}{m} \geqslant G_n(x),$$

подчиняется весьма простому распределению:

$$P(C(m, n)=j) = \frac{1}{m+n},\quad j=0,\ 1,\ldots,\ m. \qquad (9.13.9)$$

(При этом предполагается, что $m=np$, $p \geqslant 1$ — целое). При истинности альтернативы распределение величины $C(m, n)$ будет отличаться от равномерного, что позволит применить подходящий критерий согласия для проверки гипотезы однородности.

Особое замечание следует сделать в связи с тем, что эмпирическая функция распределения однозначно связана с

рангами выборочных значений и, следовательно, любой двувыборочный тест, основанный на F_m и G, может быть представлен в виде эквивалентного рангового теста. [Например, ω^2_{mn}-статистика ((9.13.8) при $\psi=1$ может быть приведена к виду:

$$\omega^2_{mn}=\frac{S_{mn}}{mn(m+n)}-\frac{4mn-1}{6(m+n)},$$

где

$$S_{mn}=m\sum_{i=1}^{m}(R_i-i)^2-n\sum_{j=1}^{n}(K_j-j)^2, \qquad (9.13.10)$$

где R_i и K_j — ранги выборочных значений $\{x_i\}$ и $\{x_j\}$ соответственно в смешанной выборке (см. Андерсон [1])]. Поэтому имеется целый класс ранговых тестов однородности (см. Гаек и Шидак [1]), которые мы рассмотрим в следующей главе вместе с другими ранговыми тестами.

Тесты на выборочных интервалах. Обобщение тестов согласия для задачи однородности возможно и по отношению к тестам, основанным на выборочных интервалах (см. § 9.4). При использовании подходящих статистик можно исключить влияние сдвига и масштаба, обеспечив тем самым решение задачи однородности при сложной гипотезе принадлежности к типу. Так, Блюменталь [2] рассмотрел в этом аспекте тесты, основанные на статистике

$$T_n(r)=S_n(r)\cdot S_n(-r),$$

где

$$S_n(r)=\sum_{i=1}^{n-1}(\Delta x_i\,\Delta y_i)^r,$$

Δx_i — выборочные интервалы первой выборки, а Δy_i — второй. При этом он обнаружил, что независимость предельного распределения статистики от гипотетического распределения наблюдается лишь асимптотически и лишь для $\frac{1}{2}\leqslant r\leqslant 1$, и рассмотрел мощностные свойства данных тестов. При больших конечных объемах выборки T_n-тест оказывается по мощности хуже колмогоровского, но лучше χ^2-теста.

К числу тестов на выборочных интервалах относится также критерий пустых блоков, подробно описанный Уилксом [1]; Кудлаев [1] показал, что критерий общего числа серий (см. Уилкс [1]) является эквивалентным критерию пустых блоков.

В заключение отметим, что вопросы решения задачи однородности при наличии связок рассматривались Нётером [3] и Фудзимото [1]; здесь мы не будем вдаваться в детали.

ГЛАВА X

О РАНГОВЫХ ТЕСТАХ

§ 10.1. ВВЕДЕНИЕ

По ряду причин статистические процедуры, основанные на рангах, привлекли большое внимание статистиков. Пристрастие к рангам иногда доходит даже до того, что их представляют наиболее важной и чуть ли не единственной основой непараметрических выводов. Например, Я. Гаек [1] дал своей книге по ранговым тестам (надо сказать, очень хорошей книге для первоначального знакомства с предметом) название «Курс непараметрической статистики». Впрочем, такое пристрастие простительно не только потому, что можно понять личную увлеченность того или иного ученого предметом своих исследований, но и потому, что ранговые процедуры действительно обладают рядом замечательных свойств. Во-первых, ранговые тесты непараметричны по своей природе: для всех гипотез, охватывающих распределения, симметричные относительно перестановок аргументов, ранговые тесты имеют фиксированный уровень значимости независимо от конкретного вида гипотетического распределения. Во-вторых, исследования ранговых тестов показали, что они в ряде случаев не уступают по мощностным свойствам параметрическим тестам при заданных распределениях гипотезы и альтернативы, и могут значительно превосходить последние при отклонениях истинных распределений от предполагаемых. Оба эти свойства ранговых тестов, по существу, являются следствиями соответствующих свойств рангов (см. гл. V): первое вытекает из равномерности распределения рангового вектора при гипотезе (см. § 5.2), второе связано с усилением статистической связи между рангами и выборочными значениями при возрастании объема выборки (см. § 5.4). Следует, однако, подчеркнуть некоторую ограниченность области применения ранговых тестов: альтернативные распределения обязательно должны быть неинвариантными к перестановкам аргументов (см. § 5.1).

§ 10.2. ОПТИМАЛЬНЫЕ И АСИМПТОТИЧЕСКИ ОПТИМАЛЬНЫЕ РАНГОВЫЕ ТЕСТЫ

Поскольку распределение P_f рангового вектора при гипотезе $f(x)$ известно (см. (5.2.3)), а при альтернативе $g(x)$ это распределение P_g дается формулой Хёфдинга (см. (5.3.4)), то, согласно фундаментальной лемме Неймана-Пирсона, наиболее мощный ранговый тест получится, если использовать в качестве тестовой статистики отношение правдоподобия (или его логарифм). Берк и Сэвидж [1] исследовали свойства такого отношения правдоподобия и показали, что оно асимптотически эквивалентно отношению правдоподобия для исходной выборки. Однако на практике вычислить P_g в явном виде обычно не удается. Эту трудность можно обойти, если ограничиться классом локально наиболее мощных ранговых тестов.

Рассмотрим, например, двувыборочную задачу, P_g для которой дается формулой (5.3.6), которую мы перепишем (обозначив $N_1+N_2=N$, $N=n$) в виде

$$P_g=\binom{N}{n}^{-1} E_f\left\{\prod_{i=1}^{N}\left[\frac{g(w_{(R_i)})}{f(w_{(R_i)})}\right]^{Z_{\alpha_i}}\right\}. \tag{10.2.1}$$

Так как $P_f=\binom{N}{n}^{-1}$, то отношение правдоподобия L отличается от (10.2.1) лишь отсутствием множителя $\binom{N}{n}^{-1}$. Предположим, что g отличается от f некоторым малым параметром θ и выполняются условия регулярности, позволяющие раскладывать L в ряд по степеням θ и проводить дифференцирование под знаком математического ожидания (подробности см., например, у Капона [1]). Тогда

$$L=1+\frac{\partial L}{\partial\theta}\cdot\theta+o(\theta), \tag{10.2.2}$$

$$\frac{\partial L}{\partial\theta}=E_f\left\{\sum_{i=1}^{N}\frac{\frac{\partial}{\partial\theta}g(w_{(R_i)})\big|_{\theta=0}}{f(w_{(R_i)})}\cdot Z_{\alpha_i}\right\}=$$

$$=\sum_{i=1}^{N}E_f\left[\frac{\partial}{\partial\theta}\ln f(w_{(R_i)})\right]\cdot Z_{\alpha_i}=\sum_{i=1}^{N}a_{N_i}\cdot Z_{\alpha_i}, \tag{10.2.3}$$

где

$$a_{N_i}=E_f\left[\frac{\partial}{\partial\theta}\ln f(w_{(R_i)})\right]=E_f\left[\frac{f'_\theta(w_{(R_i)})}{f(w_{(R_i)})}\right]. \tag{10.2.4}$$

Подставляя (10.2.3) в (10.2.2), пренебрегая членами порядка $0(\theta)$ и учитывая, что прибавление константы и умножение на константу не влияет на качество тестовой статистики, получаем, что в указанном приближении оптимальная тестовая статистика имеет вид

$$S=\sum a_{Ni} Z_{\sigma_i}. \tag{10.2.5}$$

Аналогичным образом получаются ранговые статистики для других задач. Все они имеют вид

$$S=\sum a_{Ni}\cdot C_i. \tag{10.2.6}$$

В литературе принято называть весовые коэффициенты a_{Ni} термином scores, который удобно перевести как веса (в другом переводе — метки); C_i — константами регрессии (в (10.2.5) C_i равны 1 или 0); а сами статистики вида (10.2.6) — линейными статистиками.

Конкретный вид ранговых статистик получится, если выполнить усреднение, требуемое формулой (10.2.4). Иногда сделать это бывает затруднительно, и тогда можно сделать еще один шаг аппроксимации, заменив математическое ожидание функции порядковой статистики функцией от математического ожидания порядковой статистики:

$$a_{Ni}=E_f\left[\frac{f'_\theta(x_{(R_i)})}{f(x_{(R_i)})}\right]\doteq\frac{f'_\theta\left(F^{-1}\left(\frac{R_i}{N+1}\right)\right)}{f\left(F^{-1}\left(\frac{R_i}{N+1}\right)\right)}. \tag{10.2.6}$$

Основанием для такого шага является сходимость порядковых статистик к квантилям (см. § 3.5) и предположение о гладкости усредняемой функции. В результате мы получаем линейные ранговые статистики вида

$$S=\sum\frac{f'_\theta\left(F^{-1}\left(\frac{R_i}{N+1}\right)\right)}{f\left(F^{-1}\left(\frac{R_i}{N+1}\right)\right)}\cdot C_i. \tag{10.2.7}$$

В силу сходимости порядковых статистик к квантилям, статистики (10.2.7) асимптотически эквивалентны статистикам, основанным на весах a_{Ni}, получаемых по формуле (10.2.4).

О тестах, основанных на линейных ранговых статистиках, можно сделать следующие утверждения:

1. Поскольку при нулевой гипотезе распределение рангового вектора не зависит от вида распределения выборки,

такие тесты сохраняют уровень значимости при любых распределениях выборки.

2. Так как при сближении альтернативы с гипотезой ($\theta \to 0$) разница между линейной ранговой статистикой и статистикой отношения правдоподобия исчезает, ранговые тесты являются локально наиболее мощными в области $0 \in [0, \varepsilon]$, где ε — достаточно малая величина. Другими словами, при $\theta \to 0$ и $N \to \infty$, ранговые тесты являются асимптотически оптимальными.

3. Веса a_{N_i} линейной ранговой статистики явно определяются распределением нулевой гипотезы (см. (10.2.6)) и распределением альтернативы (поскольку $g'_\theta|_{\theta=0} = f'_\theta$). Тем самым веса оказываются «привязанными» к конкретной паре (f, g). Это приводит к тому, что хотя уровень значимости не зависит от f, мощность заданного S-теста будет зависеть от того, каковы истинные f и g, и будет, конечно, максимальной при их совпадении с теми, на основе которых вычислены веса a_{N_i}. Это и является причиной существования «предпочтительных» распределений для каждого из ранговых тестов.

4. Аддитивная структура ранговых статистик обеспечивает им асимптотическую нормальность при выполнении условий центральной предельной теоремы. Для многих ранговых тестов и конкретных f и g выполнение этих условий доказано строго (см. Чернов и Сэвидж [1], Гаек и Шидак [1], Гаек [1], [2], [3] и др.). Это существенно облегчает определение приближенных критических значений для соответствующих тестов, а при асимптотической нормальности в случае истинности альтернативы — и вычисление мощностей или констант эффективности.

§ 10.3. φ-ФУНКЦИИ ГАЕКА

Для построения конкретного рангового теста необходимо нахождение весов a_{N_i}. Удобно ввести понятие φ-функции

$$\varphi(u) = \frac{f'_\theta(F^{-1}(u))}{f(F^{-1}(u))}, \quad 0 \leqslant 1 \leqslant u, \tag{10.3.1}$$

и говорить, что веса a_{N_i} порождаются φ-функцией. Хотя многие ранговые тесты были предложены различными авторами из других соображений*, введение φ-функций позволяет алгоритмически получать известные ранее и новые тесты.

* Интересно, например, отметить, что история знакового теста, который также относится к ранговым, восходит к 1710 году, когда Арбутнотт [1], анализируя с помощью знакового теста статистические данные о рождаемости мальчиков и девочек, обосновывал «мудрость провидения всевышнего».

Для каждого конкретного типа альтернативы φ-функции имеют свой вид. Например, для альтернатив сдвига

$$\varphi(u) = -\frac{f_x(F^{-1}(u))}{f(F^{-1}(u))}, \tag{10.3.2}$$

поскольку для этого типа альтернативы $f'_\theta = -f'_x$. С другой стороны, для масштабных альтернатив

$$\varphi(u) = -1 - F^{-1}(u)\frac{f'_x(F^{-1}(u))}{f(F^{-1}(u))}. \tag{10.3.3}$$

(Эту формулу легко получить, задав масштабную альтернативу в виде $g_\theta(x) = e^{-\theta} f(xe^{-\theta})$ и проведя соответствующее дифференцирование).

Конкретный вид φ-функций получается при задании семейства распределений g_θ (включающего $f = g_\theta|_{\theta=0}$). Так, для экспоненциальной альтернативы масштаба ($g_\theta = \theta e^{-\theta x}$)

$$\varphi(u) = \frac{1}{\theta_0}[1 - \ln(1-u)]; \tag{10.3.4}$$

для нормальной альтернативы сдвига

$$\left(g_\theta = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(x-\theta)^2\right)\right)$$

$$\varphi(u) = \Phi^{-1}(u), \tag{10.3.5}$$

где $\Phi^{-1}(u)$ — функция, обратная стандартной нормальной функции распределения;
для логистической альтернативы сдвига

$$\varphi(u) = 2u - 1, \tag{10.3.6}$$

и этот список можно продолжить.

Конкретная тестовая статистика получится путем вычисления a_{Ni} по формуле (10.2.4) или приближенно по формуле $\tilde{a}_{Ni} = \varphi\left(\frac{R_i}{N+1}\right)$. Так, прямое вычисление a_{Ni} для экспоненциальной альтернативы масштаба путем усреднения (10.3.4) дает тест Сэвиджа с весами

$$a_{Ni} = \sum_{j=N+1-i}^{N} \frac{1}{j}, \tag{10.3.7}$$

а асимптотически эквивалентный ему приближенный тест имеет веса

$$\tilde{a}_{Ni} = \ln\left(1 + \frac{R_i}{N+1}\right). \tag{10.3.8}$$

Аналогичным образом строятся тестовые статистики для других распределений и типов альтернатив.

§ 10.4. ОПРАВДАНИЕ КРАТКОСТИ ДАННОЙ ГЛАВЫ

Литература по ранговым тестам насчитывает сейчас многие сотни статей и большое количество монографий. Одна из них, отражающая современное состояние теории ранговых тестов и содержащая необходимые практические рекомендации и таблицы и обширную библиографию, к настоящему времени переведена на русский язык. Это книга Я. Гаека и З. Шидака [1], адресованная тем, кто хочет углубленно изучить теоретические вопросы (отметим также книгу профессора Пражского университета Я. Гаека [1], предназначенную для лиц, интересующихся прежде всего приложениями; важные результаты по мощностным свойствам многих ранговых тестов содержатся в монографии Брэдли [1]). Поэтому мы не будем в данной книге излагать другие вопросы теории, кроме уже обсужденных выше.

§ 10.5. ПРИМЕНЕНИЕ РАНГОВЫХ ТЕСТОВ К ЗАДАЧЕ ОБНАРУЖЕНИЯ СЛАБЫХ СИГНАЛОВ В ШУМАХ

Задача обнаружения слабых сигналов на фоне мешающих шумов относится к одной из важнейших прикладных задач статистики. Многие области науки, техники, практики вообще сталкиваются с необходимостью ее решения, но, пожалуй, наиболее остро эта задача стоит в радиолокации и дальней связи. Этой задаче посвящено колоссальное количество статей, книг и диссертаций, хотя в принципе это «всего лишь» задача проверки статистической гипотезы об отсутствии сигнала против альтернативы и его присутствии Эта проблема остается актуальной и сегодня и прежде всего потому, что на практике все чаще приходится сталкиваться с более сложной помеховой ситуацией и использовать более сложные реальные сигналы, чем те, которые предполагаются в развитых до сих пор моделях.

Одна из современных ветвей теории обнаружения связана с неизвестностью или нестационарностью распределений, участвующих в задаче обнаружения. Это, естественно, привлекло интерес исследователей к непараметрическим методам. Ранговые тесты при этом оказались как бы специально приспособленными к задаче обнаружения: их высокие качества проявляются именно в случае альтернатив, мало отличающихся от гипотезы, что на инженерном языке как раз

и соответствует случаю слабых сигналов или, точнее, малых отношений сигнала к шуму.

Не претендуя на полноту, дадим краткий обзор работ, посвященных использованию ранговых процедур для целей обнаружения. Можно выделить три основных группы таких работ.

Работы первой группы посвящены изложению и развитию теории ранговых тестов применительно к задаче обнаружения. Сюда следует отнести статьи Левина [1], Кушнира и Левина [1], посвященные асимптотической оптимальности ранговых тестов; работу Антоняка и Дилларда [1], в которой рассматривается последовательный вариант ранговой процедуры; статью Вольфа и Гаствирта [1], затрагивающую интересный и важный вопрос о переходе от дискретной выборки к непрерывной реализации принимаемого сигнала (при использовании знакового теста). Лаиниотис [1] предложил модификацию знакового теста, сводящуюся к введению более чем одного опорного квантиля; Тарасенко и Шуленин [3] уточнили вопрос о выборе оптимальных опорных квантилей для этой процедуры. Если опорный квантиль неизвестен, кажется естественной модификация теста, состоящая в использовании оценки этого квантиля вместо его истинного значения. Однако Гаствирт [1] показал, что такой модифицированный знаковый тест теряет непараметричность: уровень его значимости очень сильно зависит от истинного распределения. Практическая необходимость обнаружения сигнала в многоканальных системах потребовала развития теории многовыборочных ранговых тестов. Некоторые результаты в этом направлении содержатся в работах Карлайла [1], Дэли и Рашфорта [1], Столла и Курца [1], Воински и Курца [1], Зигангиров [1] предложил своеобразный многовыборочный ранговый тест, чувствительный как к альтернативе сдвига, так и к масштабной альтернативе.

Вторую группу работ образуют исследования свойств заданных ранговых тестов (прежде всего их эффективности) в различных конкретных условиях обнаружения. Вопросы обнаружения в гауссовом шуме с помощью коррелятора совпадений полярностей рассмотрены Вольфом, Томасом и Вильямсом [1], Екре [1], Хэнкоком и Лаиниотисом [1]; Миллард и Курц [1] сравнили пять тестов типа Колмогорова — Смирнова также в предположении белого гуассова шума. Случай цветного шума рассмотрен у Карлайла [1]. Многочисленные варианты возникают при конкретизации свойств полезного сигнала. Обнаружение сигналов известной формы (приводящее, естественно, к ранговым статистикам корреляционного типа) исследовалось Вольфом и Гаствиртом [1], Грёне-

вельдом [1], Томасом [1], Левиным и Кушниром [1], Кушниром и Левиным [1]. Обнаружение стохастических сигналов рассматривали Левин и Кушнир (в двух только что упомянутых статьях) и Грёневельд [1]. Специфика обнаружения амплитудно-модулированных и фазово-модулированных сигналов учтена Левиным и Рыбиным [1, 2], а также Кушниром [1]. Особенности радиолокационного обнаружения (т. е. обнаружения в одной из многих ячеек разрешения по измеряемому параметру) рассмотрены Антоняком и Диллардом [1]; интересный вариант обнаружения сигнала в одной из частотных ячеек исследовал Воински [1].

Исследования первых двух групп носят теоретический характер. По существу, они отвечают на вопросы, какими должны быть структура и алгоритм работы ранговых обнаружителей и какими свойствами будут обладать эти ранговые обнаружители, если их реализовать. Однако реализация ранговых тестов на практике наталкивается на ряд трудностей. Некоторые простые тесты (такие, как знаковый, или коррелятор полярностей) требуют порядка N операций, но имеют низкие эффективности. Более сложные и более мощные тесты требуют вычисления рангов; при этом необходимо запоминание всех выборочных значений, а число операций над ними достигает N^2. Кроме того, необходимо вычисление (или запоминание) всех весовых коэффициентов a_{N_i}, причем эти веса требуют пересчета при изменении объема выборки. Если учесть, что при обнаружении слабых сигналов объемы выборки велики (N может достигать порядка нескольких тысяч), то работа ранговых обнаружителей в реальном масштабе времени оказывается нереальной. Поэтому работы Фьюстела и Дэвиссона [1, 2, 3], посвященные поиску путей обхода этих трудностей, мы выделим в третью, самостоятельную и специфическую группу работ.

Идея Фьюстела и Дэвиссона, как и все хорошие идеи, чрезвычайно проста и состоит в том, что можно разумным образом отказаться от оптимальности теста (тем более, что каждая тестовая статистика оптимальна лишь для конкретной альтернативы), упростить статистику, несколько проиграв в эффективности, но зато существенно выиграв в простоте реализации. Для этого предлагается два (не исключающих друг друга) способа. Первый состоит в разумной аппроксимации (в простейшем случае — линеаризации) φ-функции, что даст существенное упрощение в вычислении весов. (Расчеты показывают, что потеря эффективности при этом достигает всего нескольких процентов). Второй способ направлен на сокращение числа операций при упорядочивании, опять-таки без значительных потерь в эффективности.

Это достигается путем разбиения выборки объема N на k групп наблюдений по m наблюдений в каждой ($N=km$); затем для каждой группы производится самостоятельная ранжировка (требующая m^2 операций) и вычисление тестовой статистики. Полученные k статистик складываются и результат используется для проверки гипотез. В итоге число операций имеет порядок $km^2=Nm$ вместо N^2. Оказывается, что эффективность такого «блочного» (mixed) рангового теста очень быстро возрастает с ростом m, так что обычно при $6\leqslant m\leqslant 15$ достигается от 80 до 90 процентов эффективности оптимального теста. Интуитивное обоснование этого результата состоит в том, что ранговые статистики достаточно быстро нормализуются, а для нормальных величин суммирование является оптимальным способом их комбинирования при обнаружении сдвига.

В заключение данной главы следует подчеркнуть, что все рассматриваемые выше ранговые тесты предполагают независимость выборочных значений. Появление зависимости между элементами выборки приводит к потере непараметричности всех тестов, поскольку дисперсия тестовых статистик существенно зависит от распределения выборки даже при гипотезе. Вопросы теории и практики непараметрических выводов при наличии зависимости заслуживают отдельного рассмотрения.

ЛИТЕРАТУРА

Айрлэнд, Калбэк (C. T. Ireland, S. Kullback). 1. Minimum Discrimination Information Estimation. Biometrics, 1968, **24**, N 3, 707—713.

Андерсон (Anderson T. W.). 1. On the distribution of the two-sample Cramer—von Mises criterion. AMS, 1962, **33**, 1148—1159.

Андерсон, Дарлинг (T. W. Anderson, D. A. Darling). 1. Asymptotic theory of certain goodness of fit criteria based on stochastic processes. AMS, 1952, **23**, 193—212.

Антоняк, Диллард (C. E. Antoniak, G. M. Dillard). 1. A distribution-free sequential probability-ratio test for multiple-resolution-element radars. IEEE Trans., 1968, **IT**, Nov., 822—825.

Арбутнотт (J. Arbuthnott). 1. An Argument for Divine Providence, taken from the constant Regularity observ'd in the Birth of both Sexes Phil. Trans., 1710, **27**, 186—190.

Балакшин Б. С. 1. Основы технологии машиностроения. Изд-во «Машиностроение», М., 1966.

Бартон, Дэвид (D. E. Barton, F. N. David). 1. Tests for randomness of points on a line. Biometrika, 1956, **43**, 104—112.

Бахадур (R. R. Bahadur). 1. Rates of Convergence of Estimates and Test Statistics, AMS, 1967, **38**, 303—324. 2. Stochastic Comparison of Tests, AMS, 1960, **31**, 276—295.

Белл, Доксум (C. B. Bell, K. A. Doksum). 1. „Optimal" one-sample distribution-free test and their two-sample extensions. AMS, 1966, **37**, 120—132.

Берк, Сэвидж (R. H. Berk, I. R. Savage). 1. The information in a rank-order and the stopping time of some associated SPRT's. AMS, 1968, **39**, 1661—1674.

Бернштейн С. Н 1. Теория вероятностей. Гостехиздат, 1946.

Бирнбаум (Z. W. Birnbaum). 1. Numerical tabulation of the distribution of Kolmogorov's statistics for finite sample size. Journ. Amer. Statist. Assoc., 1952, **47**, 425—440. 2. On the power of a one-sided test of fit for continuous probability functions. AMS, 1953, **24**, N 3, 484—489.

Бирнбаум, Тинги (Z. W. Birnbaum, F. H. Tingey). 1. One-sided confidence contours for probability distribution functions. AMS, 1951, 22, 592—596.

Блеквелл (D. Blackwell). 1. Conditional expectation and unbiased sequential estimation. AMS, 1947, 18, 105—110.

Блюм, Розенблатт (J. R. Blum, J. Rosenblatt). 1. On estimating quantiles. Ann. Inst. Stat. Math., 1963, 15, N 1, 45—50.

Блюменталь (S. Blumenthal). 1. Tests of fit based on partial sums of the ordered spacings. Ann. Inst. Stat. Math., 1970, 22, N 2, 261—276. 2. Contributions to sample spacings theory, I: Limit distributions of sums of ratios of spacings. AMS, 1966, 37, 4, 904—924. 3. Contributions to sample spacings theory, II: Tests of the parametric goodness-of-fit and two-sample problems, AMS, 1966, 4, 925—939. 4. Logarithms of sample spacings. SIAM J. Appl. Math., 1968, 16, 1184—1191.

Большев Л. Н., Смирнов Н. В. 1 Таблицы математической статистики. Изд-во «Наука», М., 1965.

Боровков А. А. 1. К задаче о двух выборках. Изв. АН СССР, серия матем., 1962, 26, 605—624.

Бохнер (S. Bochner). 1. Harmonic Analysis and the Theory of Probability. Univ. of Calif. Press., 1955.

Де Брейн Н. Г. 1. Асимптотические методы в анализе. Изд-во «ИЛ», М., 1961.

Бриллюэн Л. 1. Наука и теория информации. ГИФМЛ, М., 1960 2. Научная неопределенность и информация. Изд-во «Наука», 1965.

Брэдли (J. V. Bradley). 1. Distribution-free statistical tests. Prentice-Hall, 1968.

Бурр (E. F. Burre). 1. Distribution of the two-sample Cramer — von Mises criterion for small equal samples. AMS, 1963, 34, 95—101.

Бхаттачарья (P. K. Bhattacharya). 1. Estimation of probability density fanction and its derivatives. Sankhya, 1967, A29, N 4, 373—382.

Бхаттачарья, Рушас (P. K. Bhattacharya, Roussas). 1. Estimation of a certain functional of a probability density function. Skand. aktuarietidskr., 1969, N 3—4, 201—206.

Вайсс (L. Weiss). 1. Confidence intervals of preassigned length for quantiles of unimodal populations. Naval Res. Logistics Quarterly, 1960, v. 7, 251—256. 2. The limiting joint distribution of the largest and smallest sample spacings. AMS, 1959, 30, 590—593. 3. On the asymptotic power of Crame'r — von Mises tests of fit. Ann. Inst. Stat. Math., 1966, 18, N 2, 149—153. 4. Tests of fit in the presence of nuisance location and scale parameters. AMS, 1957, 28, 1016—1020. 5. The asymptoticpower of certain tests of fit based on sample spacings. AMS, 1957, 28, 783—786. 6. On asymptotic sampling theory for distributions approaching the uniform distribution. Z. Warscheinlichkeitstheorie Verw. Geb., 1965, 4, 217—221. 7. Review of paper by Chibisov. Math. Rev., 1962, 24, N. A1776.

Вальд (A. Wald). 1. An extension of Wilks' method for setting tolerace limits. AMS, 1943, 14, 45—55.

Ван Иден К. (C. Van Eeden). 1. The relation between Pitman's asymptotic relative efficiency of two tests and the correlation coefficient between their tests statistics. AMS, 1963, 34, N 4, 1442—1450.

Ван-Цвет (W. R. Van Zwet). 1. „Convex Transformations of Random Variables", Amsterdam: Math. Centre., 1964.

Вильямс (C. A. Williams). 1. On the choice of the number and width of classes for the chi-square-tests of goodness of fit. J. Am. Stat. Ass., 1950, 45, 77—86.

Воински (M. N. Woinsky). 1. Nonparametric detection using spectral data. IEEE Ttrans, 1972, IT, 18, N 1, 110—118.

Вольф, Гаствирт (S. S. Wolf, J. L. Gastwirth). 1. The effect of autoregressive dependence on a nonparametric test. IEEE Trans., 1967, IT, April, 311—313.

Вольф, Томас, Вильямс (E. S. Wolf, J. B. Thomas, T. R. Williams). 1. The polarity-coincidence correlator: a nonparametric detection device. IRE Trans., 1962, IT, 5—9.

Вольфовиц (J. Wolfowitz). 1. Additive Partition Functions and a Class of Statistical Hypotheses. AMS, 1942, 13, 247—279. 2. Generalization of the theorem of Glivenko-Cantelli. AMS, 1954, vol. 25, 131—138.

Вулвертон, Вагнер (C. T. Wolverton, T. J. Wagner). 1. Recoursive estimates of probability densities. IEEE Trans, 1969, SSC-5, N. 3, 246—247.

Гаек Я. (J. Hajek). 1. „Course in Nonparametric Statistics", Akademic Press, 1970. 2. Some extention of the Wald-Wolfowitz-Noether theorem. AMS, 1961, 32, 506—523. 3. Asymptotic normality of simple linear rank statistics under alternatives. AMS, 1968, 39, N. 2, 325—346. 4. Miscellaneous problems of rank test theory. „Nonparam. Techniques in Stat. Inference", Univ. Press., Cambrige, 1970, 3—19.

Гаек Я., Шидак З. Теория ранговых критериев. «Наука», М., 1971.

Гаствирт (J. L. Gastwirth). 1. On the sign test of symmetry. J. Am. Stat. Soc., 1971, 66, N. 336, 821—823.

Геберт (J. R. Gebert). 1. A Power Study of Kimball's statistics. Statistische Hefte, 9, N. 4, 269—273.

Геберт, Кейл (J. R. Gebert, B. K. Kale). 1. Goodness of fit tests based on discriminatory information. Statistische Hefte, 1969.

Гельцер (J. W. Gelzer). 1. The asymptotic relative efficiency of several goodnees of fit tests. M. S. Thesis, Univ. of Washington (1962).

Гельцер, Пайк (J. Gelzer, R. Pyke). 1. The asymptotic relative efficiency of goodness-of-fit tests against scalar alternatives. J. Am. Stat. Assoc., 1965, N. 60, 410—419.

Гихман И. И. 1. Некоторые замечания к критерию согласия А. Н. Колмогорова. ДАН СССР, 1953, XCI, № 4, 715—718. 2. Об эмпирической функции распределения в случае группировки данных. ДАН СССР, 1952, XXXII, № 6, 837—840.

Гихман И. И., Гнеденко Б. В., Смирнов Н. В. 1. Непара-

метрические методы статистики. Труды III Всесоюзного математ. съезда, т. III, М., Изд-во АН СССР, 1956, 320—334.

Гихман И. И., Скороход А. Б. 1. Введение в теорию случайных процессов. Изд-во «Наука», М., 1965.

Гливенко В. 1. Sulla determinazione empirica della leggi di probabilita. Giorn. dell' L'instit. degli att. 1933, vol. 4, 92—99.

Гнеденко Б. В. 1. Sur la distribution limite du terme maximum d'une se'rie aleatoire. Ann. Math., 1943, 44, 423, 453. 2. Предельные теоремы для максимального члена вариационного ряда. ДАН СССР, 1941, **32**, 7—9. 3. Курс теории вероятностей. «Наука», 1969.

Гнеденко Б. В., Королюк В. С. 1. О максимальном расхождении двух эмпирических распределений. ДАН СССР, XXX, 1951, № 4, 525—528.

Гнеденко Б. В., Михалевич В. С. 1. Две теоремы о поведении эмпирических функций распределения. ДАН СССР, 1952, XXXV, № 1, 25—27. 2. О распределении числа выходов одной эмпирической функции над другой. ДАН СССР, 1952, XXXII, № 6, 841—843.

Гнеденко Б. В., Рвачева Е. Л. 1. Об одной задаче сравнения двух эмпирических распределений. ДАН СССР, 1952, т. XXXII, № 4, 513—516.

Греневельд (R. A. Groeneveld). 1. A nonparametric rank correlation method for detecting signals in additive noise. IEEE Trans., 1967, IT, Apr, 315—316.

Греневельд, Хинч (R. A. Groeneveld, M. J. Hinch). 1 The efficiency of level tests in the detection of weak stochastic signals. IEEE Trans., 1969, IT, July, 502—504.

Гринвуд (M. Green'wood). 1. The statistical study of infection diseases. J. Roy, Stat. Soc., 1946, 109, 85—110.

Гумбель (E. J. Gumbel). 1. Les valeurs extre'mes des distributions statistiqnes. Ann. Inst. H.; Poincare', 1934, 5, 115. 2. On the reliability of the classical x^2-test. AMS, 1943, **14**, 253—263.

Дарлинг Д. (D. A. Darling). 1. The Kolmogorov—Smirnov, Cramе'r—von Mises Tests. AMS, 1957, 28, 823—838. 2. On a class of problems related to the random division of an interval. AMS, 1953, **24**, 239—253. 3. The Cramer—Smirnov Test in the parametric case. AMS, 1955, 26, N. 1, 1—20.

Джонсон Н. (N. L. Johnson). 1. Systems of freqnency curves generated by methods of translation. Biometrica, 1949, **36**, 149.

Дмитриев Ю. Г., Тарасенко Ф. П. 1. О статистическом оценивании функционалов от плотности вероятностей ТВиЕП, 1974, XIX, № 2, 404—409.

Донскер (M. D. Donsker). 1. Justification and extension of Doob's heuristic approack to the Kolmogorov—Smirnov theorems. AMS, 1952, 23, 277—281.

Дуб (J. L. Doob). 1. Heuristic approach to the Kolmogorov — Smirnov limit theorems. AMS, 1949, **20**, 393—403.

Дурбин (J. Durbin). 1. Some methods of constructing exact tests. Biometrika, 1965, **48**, 41—55.

Дэвид (H. A. David). 1. „Order Statistics", Wiley, N-Y, 1970

Дэвиссон (L. D. Davisson). 1. Convergence probability bounds for stochastic approximation.

Дэли, Рашфорт (R. F. Daly, C. K. Rushforth). 1. Nonparametric detection of a signal of known form in additive noise. IEEE Trans., IT, 1965, JAN., 70—76.

Екре (H. Ekre). 1. Polarity coincidence correlation detection of a weak noise source. IEEE, Trans., 1963, IT-9, N. 1, 18—23.

Епанечников В. А. 1. Непараметрическая оценка многомерной плотности вероятности. Т. вер. и ее прим., 1969, XIV, № 1, 156—160.

Жирова Л. П., Тарасенко Ф. П., Шуленин В. П. 1. Об экспериментальном сравнении тестов согласия. Труды СФТИ, 1972, вып. 63.

Залер (W. Sahler). 1. A Survey on distribution-free statistics based on distuncss between distribution functions. Metrik'a, 1968, 13, N 2—3, 149—169.

Зигангиров К. Ш. 1. Один непараметрический критерий сравнения выборок, ППИ, 1965, 1, № 3, 118—121.

Карлайл (J. W. Carlyle). 1. Nonparametric methods in detection theory. „Communication Theory", ed by Balakrishnan, ch. 8.

Капон (J. Kapon). 1. Asymptotic efficiency of certain LMPRT's. AMS, 1961, 32, 88—100. 2. On the asymptotic efficiency of the Kolmogorov—Smirnov test. J. Amer. Statist. Assoc., 1965, 60, 843—853.

Кац, Кифер, Вольфовиц (M. Kac, J. Kiefer, J. Wolfowitz). 1. On tests of normality and other goodness of fit tests based on distance method. AMS, 1955, 26, 189—211.

Кашньяп, Блэйдон (R. L. Kashnyap, C. C. Blaydon). 1. Estimation of Probability density and distribution function. Trans IEEE, 1968, IT-14, N. 4, 549—556. 2. Recovery of functions from noisy measurements. Proc IEEE, 1966, 54, 1127—1129.

Квайд (Quade D.). 1. The asymtotic power of the Kolmogorov tests of goodness of fit Inst. Stat. Mimeo. Series, N. 243, Univ. of Nokth Carolina, Chapel Hill, 1959.

Кейл (Kale B. K.). 1. Unified derivation of tests of goodness of fit based on spacings. Sankhya, 1969, 1, 43—48. 2. Tests of goodness of fit based on discriminatory information. Abstract, AMS, 1965, 36, 1601.

Кейл, Годамбе (B. K. Kale, V. P. Godambe). 1. A test of goodness of fit. Statictishe Hefte, 1967, 8, 165—172.

Кендалл, Стьюарт (M. G. Kendall, A. Stuart). 1. „The Advanced Theory, of Statistics", vol. 2, Charles Griffin, London, 1962.

Кендалл (M. G. Kendall). 1. „Rank Correlation Methods" Charles Griffin, London, 1953.

Кимбалл (B. F. Kimball). 1. Some basic theorems for developing tests of fit for the case of non-parametric probability distribution function. I. AMS, 1947, 18, 540—548.

Клотц (J. Klotz). 1. Asymptotic efficiency of the tuvo sample Kolmogorov — Smirnov test. J. Amer Stat. Assoc., 1967, 62, N. 319, 932—938.

Ковер (T. Cover). 1. Сообщение на II Международном симпозиуме по теории информации, Цахкадзор, 1971.

Кокран (W. G. Cochran) 1. The x^2-test of goodness of fit. AMS, 1952, 23, 315—345.

Колмогоров А. Н. 1. Sulla determinazionne empirica di une legge di distribuzionne. Giorn. Instit. Ital. att., 1933, 4, 83—91.

Коваленко И. Н., Филиппова А. А. 1. Теория вероятностей и математическая статистика. М, «Высш. школа», 1973.

Колмогоров А. Н., Фомин С. В. 1. Элементы теории функций и функционального анализа. «Наука», 1968.

Комо (J. J. Komo). 1. An application of non-parametric statistics of the Kolmogorov — Smirnov type to the measurement selection problem. Proc. 3rd. Hawaii Int. Conf. on System Sci., 1970, West. Period.

Конюховский В. В. 1. Критерии согласия, однородности и независимости. Изд-во МГУ, 1970.

Корвальхо (P. De oliveira Corvalho). 1. On the distribution of the Kolmogorov — Smirnov D-statistic AMS, 1959, 30, 173—176.

Королюк В. С. 1. Асимптотический анализ распределений максимальных уклонений в схеме Бернулли. Теория вероятностей и ее применения, 1959, 4, 183—186.

Крамер (H. Cramér). 1. Математические методы статистики. ИЛ, М., 1948.

Криц, Талако (T. A. Kritz, J. V. Talako). 1. Eqnivalence of the Maximum Likelihood Estimator to a Minimum Entropy Estimator. Trabajos de estadistica y de investigacion operativa, 1968, 19, N. 1—2, 56—65.

Кронмал, Тартер (R. Kronmal, M. Tarter). 1. The estimation of probability densities and cumulatives by Fourier series methods J. Am. Stat. Assoc., 1968, 63, 925—952.

Крэйг (A. T. Craig). 1 On the distributions of certain statistics, Amer. Journ. Math., 1932, vol. 54, 353—366.

Кудлаев Э. М. 1. О некотором классе непараметрических статистик. Труды СФТИ, 1972, вып 63.

Кумар (M. C. Kumar). 1. A recurrence relation for distribution function of order statistics from bivariate distribution. J. Amer. Stat. Assoc., 1969, 64, N. 326, 600—601.

Купер (M. C. Kuiper). 1. Tests concerning random points on a cirde. Proc. Koninkl. Nederl. Akad. van Wett. 1960; Ser A, 63, NI, 38—47

Кушнир А. Ф. 1. Ранговые алгоритмы последетекторного обнаружения сигналов. Радиотехника, 1971, 26, № 4, 33—37.

Кушнир А. Ф., Левин Б. Р. 1. Оптимальные алгоритмы обнаружения сигналов в шумах. Р. и Э., 1969, № 2, 249—258.

Лаиниотис (D. G. Lainiotis). 1. Generalized distribution-free coincidence detection procedure. IEEE Intern Conv REC., 1965, pt. 7, 214—219.

Левин Б. Р. 1. Оптимальные алгоритмы обнаружения сигналов, устойчивые к изменению априорных данных «Изв. вузов. Радиотехника», 1970, XIII, № 2, 110—121.

Левин Б Р, Куншир А Ф 1 Асимптотически оптимальные алгоритмы обнаружения и различения сигналов на фоне помех РиЭ, 1969, № 2, 249—258

Левин Б Р, Рыбин А К 1 Непараметрические амплитудные и фазовые методы обнаружения сигналов I Изв АН СССР, Техн киб, 1969, № 5, 110—118 2 Непараметрические амплитудные и фазовые методы обнаружения сигналов II Изв АН СССР, Техн киб, 1970, № 1, 153—160

Левин М А, Завалий Ю И, Захваткина С А 1 Определение погрешностей ориентации методом Монте Карло при сборке деталей «Механизация и автоматизация производства» 1971, № 4

Леман Э (E L Lehmann) 1 Проверка статистических гипотез Изд во «Наука», М, 1964 2 Consistency and unbiasedness of certain nonparametric tests AMS, 1951, 22, 165—179

Леман и Шеффе (E L Lehmann, H Sheffe) 1 Completeness, similar regions, and unbiased estimation, I Sankhya, 1950, 10, 305—340

Литтифорс (H W Lilliefors) 1 On the Kolmogorov — Smirnov test for normality with mean and variance unknown J Am Stat Ass, 1967, 62, N 318, 399—402 2 On the Kolmogorov — Smirnov test for the exponential distribution with mean unknown J Am Stat Ass, 1969, 64, 387—389

Лойнс (R M Loynes) 1 Some aspects of the estimation of quantiles J Roy Stat Soc, 1966, B28, N 3, 497—512

Лофтсгарден, Квезенберри (D O Loftsgaarden, C P Quesenberry) 1 A nonparametric estimate of a multivariate density function AMS, 1965 1049—1051

Льюис (Lewis P A W) 1 Distribution of the Anderson — Darling statistic AMS, 1961, 32, 1118—1123

Малмквист (S Malmquist) 1 On certain confidence contours for distribution function AMS, 1954, 25, 523—533

Малов А Н 1 Механизация и автоматизация сборочных работ в приборостроении Изд во «Машиностроение», М, 1964

Манн, Вальд (H B Mann, A Wald) 1 On the choice of the number of class of intervals in the application of the chi square test AMS, 1942, 13, 306—317

Мизес (Von Mises R V) 1 „Mathematical theory of probability and statistics", N Y, Acad Press, 1964

Миллард, Курц (J B Millard, L Kurz) 1 The Kolmogorov — Smirnov tests in signal detection IEEE Trans, 1967, IT, apr, 341—344

Митропольский А К 1 Техника статистических вычислений Изд во «Наука», 1971

Моран (P A P Moran) 1 The random division of an interval Pt I, Journ Roy Stat Soc, 1947, Suppl, 9, 92—98 2 The random division of an interval Pt II, Journ Roy Stat Soc, 1951, B13, 147—150

Мостеллер (F Mosteller) 1 On some useful „inefficient" statistics AMS, 1946, 17, 377—407.

Мулдон (J G Mauldon) 1 Random division of an interval Proc Camb Phil Soc, 1951, 47, 331—336

Мэрфи (V K Murthy) 1 Estimation of probability density AMS, 1965, 36, N 3, 1027—1031 2 Nonparametric estimation of multivariate densities with applications

Мэсси (Massey F T) 1 A note on the power of nonparametric tests AMS, 1950, 20, 440—442 2 The Kolmogorov — Smirnov test for goodness of fit J Am, Stat Ass, 1951, 46, 68—78

Надарая Э А 1 О непараметрических оценках плотности вероятности и регрессии Т вер и ее прим, 1965, 199—203 2 Труды ВЦ АН Груз ССР, 1966, VII, № 1, 35—42

Нетер (G E Noether) 1 „Elements of Nonparametric Statistics" Wiley, N—Y, 1967 2 On a Theorem of Pitman AMS, 1955, 26, 64—68 3 Note on the Kolmogorov statistic in the discrete case Metrika, 1963, 7, N 2, 115—116

Оуэн (D B Owen) 1 «Сборник статистических таблиц» Изд во «Наука», М, 1966

Пайк (R Pyke) 1 Spacing J Roy Stat Soc, 1965, B27, 395—436 2 The supremum and infimum of the Poisson process AMS, 1959, 30, 568—576

Паттер (J. Putter) 1 The treatment of ties in some nonparametric tests AMS, 1955, 26, 368—386

Парзен Э (E Parzen) 1 On estimation of a probability density function and mode AMS, 1962, 33, 1065—1076

Пирсон (E S Pearson) 1 The probability integral transformation for testing goodness-of-fit and combining independent tests of significance Biometrica, 1938, 30, 134—148

Пирсон, Стефенс (E S Pearson, M A Stephens) 1 The goodness of fit tests based on W and U ² Biometrika, 1962, 49, 397—402

Пирсон К (K Pearson) 1 On the criterion that a given system of deviations from the probable in the case of correlated systems of variables is such that it can be reasonably supposed to have arisen from random sampling Phil Mag, 1900, vol 50, ser 5, 157—172

Поршан, Пайк (F Porshan, R Pyke) 1 Tests for monotone failure rate Proc 5th Berkeley Symp Math Stat, ç Prob, 1965

Пьерри (P A Pierre) 1 Singular Non-Gaussian Measures in Detection and Estimation Theory IEEE Trans, IT, 1969, 15, N 2, 266—272

Рамачандрамурти (P V Ramachandramurty) 1 On the Pitman efficiency of one-sided Kolmogorov and Smirnov tests for normal alternatives AMS, 1966, 37, N 4, 940—944

Рао М (M M Rao) 1 Theory of Order Statistics Math Annalen, 1963, 147, 298—312

Рао С Р (S R Rao) 1 Линейные статистические методы и их применения Изд во «Наука», М, 1968

Реньи (A Rényi) 1 On the theory of order statistics Acta Math Hung, 1953, 4, 191—232 2 „Probability Theory" Academia Kiado', Budapest, 1970

Розенблатт (M Rosenblatt) 1 Remarks on some nonparametric estimates of a density functions AMS, 1956, 27, 832—837 2 Density

estimates and Markov sequencies. „Nonparametric Techniques in Statistical Inference", ed. by. M. L. Puri, Cambridge Univ. Press, 1970, 199—210. 3. Limit theorem associated vith variants of the von mises statistic. AMS, 1952, 23, 617—623.

Руист (E. Ruist). 1. Comparison of Tests ofr Nonparametric Hypotheses. Arkiv för Mathem., 1955, B3, N. 9, 133—163.

Сархан, Гринберг (A. E. Sarhan, B. G. Greenberg). 1. „Contributions to Order Statistics". Ed. by A. E. Sarhan 8c B. G. Greenberg. Wiley, N—Y, 1962. Русский перевод: «Введение в теорию порядковых статистик», под ред. А. Я. Боярского, «Статистика», М., 1970.

Себестьен, Эдл (G. Sebestyen, J. Edle). 1. An algorithm for nonparametric pattern recognition. IEEE Trans., EC, 1966, 15, N. 6, 908—915. Перевод в «Экспресс-информации», Техн. Киб., 1967, № 19.

Сетураман, Рао (J. Sethuraman, J. S. Rao). 1. Pitman efficiences of tests based on spacings. Nonparametric Techn. in Statist. Inference", Univ. Press, Cambridge, 1970.

Симахин В. А., Тарасенко Ф. П., Шуленин В. П. 1. О непараметрическом оценивании одного класса функционалов. Труды V конф. по Т. кодир. и перед. информ. Москва — Горький, 1972, вып. 1, 112—116.

Смирнов Н. В. 1. Sur les écarts fe la courbe distribution empirique. Мат. сборник, 1939, 6 (48), 3—26. 2. Table for estimating the goodness of fit of empirical distribution. AMS, 1948, 18, 279—281. 3. О критерии Крамера — фон Мизеса. УМН, 1949, 4, 196—197. 4. Асимптотическая мощность некоторых непараметрических критериев. Труды Всес. Сов. по теории вер. и стат., Ереван, 1960, 98—100. 5. Оценка расхождения между эмпирическими кривыми распределения. Бюлл. Моск. ун-та, 1939, в. 2, 3—14. 6. Приближение законов распределения. УМН, 1944, вып. X, 179—206.

Соболев С. Л., Китов А. И., Ляпунов А. А. 1. Основные черты кибернетики. Вопросы философии, 1955, № 4.

Сринивасан (R. Srinivasan). 1. Anapproach to testing the goodness of fit of incompletely specified distributions. Biometrika, 1970, 57, N. 3, 605—611.

Столл, Курц (E. D. Stoll, L. Kurz). 1. A contribution fo the multiple-polarity coincidence detection schemes. IEEE Trans., 1968, Com-16, 676—682.

Стьюарт (A. Stuart). 1. The correlation between variate-values and ranks in sample from a continuous distribution. Brit. J. Psych. (Stat. Section), 1954, 7, 37.

Сухатме (P. V. Suhatme). 1. On the analysis of k samples from exponential popnlatiens with especial reference to the problem of random interval. Stat. Res. Mem., 1, 94—112.

Сэвидж (I. R. Suvage). 1. Nonparametric Statistics: a Personal Review. Sankhya, 1969, 31A, Pt. 2, 107—144.

Тарасенко Ф. П. 1. К определению понятия «информация» в кибернетике. Вопросы философии, 1963, № 4, 76—84. 2. Некоторые вопросы непараметрической статистики. Научная сессия, посвященная Дню радио и Дню связиста, Тезисы докладов, М., 1971, 6—8. 3. On evalution of an

unknown probability density fuction the direct estimation of entropy from in dependent observations of a continuous random variable and the distribution-free entropy test of goodness-of-fit. Proc. IEEE, 1968, 56, N. 1. 4. Введение в курс теории информации. Изд-во ТГУ, 1963. 5. Некоторые вопросы теории измерений с точки зрения теории информации. Труды СФТИ, 1958, вып. 37, 212—218. 6. Свойства выборочных блоков как основа непараметрических процедур. Радиотехника, 1971, № 4, 45—50. 7. О свойствах линейных статистик структуры D. II Международн. симпозиум по Т. инф., Цахкадзор, 1971. 8. (Редактор) «Непараметрическое оценивание интегральных функционалов по стационарным выборкам». Изд-во ТГУ, 1974.

Тарасенко Ф. П., Шуленин В. П. 1. О статистической связи между наблюдением и его рангом. Труды СФТИ, вып. 62, 1971. 2. Об одном простом методе сравнения мощностей тестов согласия структуры. Труды СФТИ, вып. 63, 1972. 3. О некоторых свойствах одной непараметрической процедуры обнаружения сигналов в шумах. Радиотехника, 1971, № 4, 42—45. 4. К использованию свойств выборочных интервалов. «Матер. III научной конф. по мат. и мех.». Изд-во ТГУ, 1973, 120—121.

Толстов Г. П. 1. «Ряды Фурье». ГИФМЛ, М., 1963.

Томас (Thomas). 1. Непараметрические методы обнаружения сигналов ТИИЭИР, 1970, 58, 5, 23—31.

Тюки (J. W. Tuckey). 1. Nonparametric estimation. II. AMS, 1947, 18, 529—539.

Хатри (C. G. Khatri). 1. Distribution of order statistics for discrete case. Ann. Inst. Statist. Math., Tokyo, 1962, 14, 167—171.

Хемельрик (J. Hemelrijk). 1. A theorem of the Sign Test when ties are persent. Proc. Koninkl Nederl Acad van Wetensch., 1952, 55A, 326—332.

Хёфдинг (W. Hoeffding). 1. Optimum nonparametric tests Proc. 2d Berkeley Symposium. University of California Press, 1951. 2. A class of statistics with asymptotically normal distributions. AMS, 1948, 19, 293—325.

Ходжес, Леман (J. L. Hodges, E. L. Lehmann). 1. The Efficiency of Some Nonparametric Competitors of the t-Test. AMS, 1956, 27, 324—335. 2. Estimates of location based on rank tests. AMS, 1963, 34, 598—611.

Худсон. 1. Статистика для физиков. «Мир», 1967.

Хэнкок, Лаиниотис (J. C. Hancock, D. G. Lainiotis). 1. On learning and distribution-free coincidence detection procedure. IEEE Trans, 1965, IT, 272—280.

Чанда (K. C. Chanda). 1. Some aspects of estimation of probability density function. Calcutta Stat. Assoc. Bull., 1966, 153—163.

Чернов, Сэвидж (H. Chernoff, I. R. Savage). 1. Asymptotic Normality and Efficiecy of Certain Nonparametric Test Statistics. AMS, 1958, 29, 972—994.

Честер (P. R. Chester). 1. Adaptive nonparametric classification. Technometrics, 1969, 11, № 4, 775—792. Перевод в «Экспресс-информации», Техн. киб., 1970, № 18.

Чибисов Д. М. 1. Об асимптотической мощности и эффективности критерия. ДАН СССР, 1961, 138, № 2, 322—325. О критериях согласия, основанных на выборочных промежутках. ТВиЕП, 1961, 6, № 3, 354—358.

Уилкс (S. Wilks). 1. Математическая статистика, Изд-во «Наука», М., 1967.

Уолш (J. E. Walsh). 1. „Handbook of Nonparametric Statistics", Van Nostrand, Vol. 1—1962, Vol. 2—1965, Vol. 3—1969. 2. Nonparametric mean estimation of percentage points and density function vales. AMS, 1957, 8, N. 3, 167—180.

Уотсон, Лидбеттер (G. S. Watson, M. R. Leadbetter). 1. On the estimation of the probability density. AMS, 1963, 34, 261—270

Файн (T. Fine). 1. On the Hodges and Lehmann shift estimator in the two sample problem. AMS, 1966, 37, N. 6, 1814—1818.

Филиппова А. А. 1. Теорема Мизеса о предельном поведении функционалов от эмпирической функции распределения и ее статистические применения. ТВиЕП, 1962, 7, № 1, 26—60.

Фишер (R. A. Fisher). 1. The conditions under which chi square measures the discrepancy between observation and hypothesis. J. Roy. Stat. Soc., 1924, vol. 87, 442—450.

Фрейзер (D. A. S. Fraser). 1. „Nonparametric Methods in Statistics", Wiley, N-Y, 1957.

Фудзимато (Hirosi Hudimoto). 1. On a two-sample nonparametric test in the case that ties are present. ANN. Inst. Math., 1959, 11, N. 2, 113—120.

Фьюстел, Дэвиссон (E. A. Feustal, L. D. Davisson). 1. The asymptotic relative efficiency of mixed statistical tests. IEEE Trans, 1967, IT-13, N. 2, 247—255. 2. Nonparametric detection by rank statistics-a unified view. IEEE Intern. Conv. Rec., 1967, 34—39. 3. On efficiency of mixed locally-most-powerful one-and two-sample rank tests. IEEE Trans., 1968, IT, sept, 776—778.

Хан Г., Шапиро С. 1. Статистические модели в инженерных задачах, изд-во «Мир», 1969. 2. Реферат. РЖ «Математика», 1973, 8В, 168.

Чэпмен (D. G. Chapman). 1. A comparative study of several goodness-of-fit tests. AMS, 1958, 29, 655—673.

Шафер, Финкельштейн, Колинз (R. E. Shafer, J. M. Finkelstein, Collins). 1. On a goodness of fit test for exponential distribution with mean unknown. Biometrica, 1972, 59, N. 1, 222—224.

Шерман (B. Sherman). 1. A random variable related to the spacing of sample values. AMS, 1950, 21, 339—361.

Штейнхаус (H. Steinhaus). 1. Sur les fonctions independentes, VIII, Studia Math, 1949, vol. 11, 133—144.

Шуленин В. П. 1. Об одном свойстве выборочных интервалов из равномерного распределения. Тр. СФТИ, 1971, вып. 62, 229—231.

Шустер (E. F. Schuster). 1. Estimation of a probability density functions and its derivatives. AMS; 1969, 40, N. 4, 1187—1195.

Цыпкин Я. З. 1. «Основы теории обучающихся систем», «Наука», М., 1970.

Феликс Петрович ТАРАСЕНКО

НЕПАРАМЕТРИЧЕСКАЯ СТАТИСТИКА

Томск, изд. ТГУ, 1976 г., 294 с.

Редактор **М. И. Сваровская.**
Технический редактор **Р. М. Подгорбунская.**
Корректоры **В. Г. Лихачева, Н. И. Викторенко.**