На правах рукописи

ЦИТОВИЧ ФЕДОР ИВАНОВИЧ

Субоптимальные последовательные статистические решения, основанные на независимых наблюдениях

Специальность 05.13.17 — Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико–математических наук

Москва — 2011 г.

Работа выполнена в Учреждении Российской академии наук
Институте проблем передачи информации им. А.А. Харкевича РАН

Научный руководитель: доктор физико–математических наук, профессор
Владимир Вячеславович Вьюгин

Официальные оппоненты: доктор физико–математических наук, доцент
Райгородский Андрей Михайлович
доктор физико–математических наук
Ройтберг Михаил Абрамович

Ведущая организация: Математический институт им. В.А. Стеклова РАН

Защита состоится "____"____________ 2011 г. в ________ час. на заседании диссертационного совета Д 002.077.01 при Институте проблем передачи информации им. А.А. Харкевича РАН по адресу: 127994, Москва, пер. Большой Каретный, дом 19, стр. 1, ауд. 615.

С диссертацией можно ознакомиться в библиотеке ИППИ РАН.

Автореферат разослан ____________ 2011 г.

Ученый секретарь диссертационного
совета Д 002.077.01
доктор физико–математических наук ____________И.И. Цитович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Задача обеспечения надежных статистических выводов всегда рассматривалась как важнейшая задача в обработке данных наблюдений. В последнее время наиболее важными становятся такие задачи при разработке автоматизированных систем управления технологическими процессами и транспортными средствами. При этом необходимо иметь ввиду, что поступающие в режиме реального времени данные могут быть подвергнуты воздействию неконтролируемых внешних причин, которые, в свою очередь, могут приводить и к потере части данных или к существенному искажению данных. Учет особенностей формирования доступных данных приводит к тому, что такие особенности могут быть адекватно описаны с помощью лишь непараметрических множеств распределений.

Наиболее сложными для исследований являются задачи, в которых требуется получать решения с гарантированным качеством. Это принципиально отличает рассматриваемую постановку задачи от байесовского подхода, который позволяет сохранить предложенную Вальдом конструкцию решающего правила, но при этом критерием качества решающего правила является средняя вероятность ошибки. Поэтому актуальной является задача построения решающих правил, гарантирующих, что что вероятность ошибки для любого допустимого распределения наблюдений не превосходит заданный уровень.

Оказалось, что решение задачи в такой подстановке существенно сложнее, поскольку в этом случае необходимо исследовать поведение целого семейства отношений правдоподобия, которое, как указано выше, является бесконечномерным.

Понятно, что решение должно приниматься за минимально короткое время, что особенно важно при управлении процессами в режиме реального времени. Поэтому строящиеся правила должны иметь свойства, близкие к оптимальным по средней продолжительности необходимых наблюдений. Таким образом возникает проблема построения гарантийного статистического решения задачи последовательной проверки непараметрических гипотез, требующего минимального количества наблюдений при любом возможном неблагоприятном распределении.

Последовательная проверка гипотез стала развиваться с 1940-ых годов. Первые результаты, легшие в основу теории последовательных статистических решений, приведены в работах A. Wald, H. Chernoff и др. Большое количество работ посвящено асимптотической теории последовательной проверки гипотез (G. Schwarz, J. Kiefer, J. Sacks, G. Lorden, А.А. Новиков, М.Б. Малютов, И.И. Цитович и др.). М.Б. Малютовым и И.И. Цитовичем была предложена асимптотически оптимальная последовательная стратегия проверки непараметрических гипотез при наличии управления наблюдениями и зоны безразличия с гарантийным решающим правилом. В основе этой стратегии лежит тест Вальда, но для ее реализации используется на первом этапе построение состоятельной оценки истинного распределения. С практической точки зрения задача построения состоятельной оценки распределения может оказаться более сложной, чем проверка гипотез, особенно в случае, когда распределения близки, поэтому практическая реализация такой стратегии может оказаться неприемлемой.

Таким образом, для обеспечения надежности статистических выводов актуальной является задача построения гарантийного статистического решения задачи последовательной проверки непараметрических гипотез, позволяющая минимизировать среднее значение необходимого количества наблюдений при любом возможном неблагоприятном распределении, но в предположении, что основные характеристики распределений известны лишь с определенной точностью. Такая постановка задачи позволяет избежать трудностей, связанных с построением состоятельной оценки распределения наблюдений.

Цель работы. Разработать непараметрические статистические модели наблюдений в зависимости от априорной информации о возможных погрешностях в распределении наблюдений. Построить субоптимальные (близкие к оптимальным) последовательные гарантийные процедуры проверки гипотез для таких моделей и исследовать их свойства.

Методы исследования. В работе использованы методы теории вероятностей, математической статистики, теории оптимизации, компьютерное моделирование.

Научная новизна работы. Предложена постановка задачи проверки статистических гипотез, в которой можно обеспечить гарантийные статистические решения, близкие к оптимальным, не используя предварительную оценку распределения наблюдений. Полученные неасимптотические оценки функции риска процедуры позволяют проводить анализ статистических моделей задачи и выбирать ту из них, которая обеспечивает наиболее эффективное решение задачи проверки гипотез при заданных априорных сведениях о возможных ошибках в наблюдениях.

Личный вклад. Все научные результаты, приведённые в диссертационной работе, получены автором лично.

Практическая ценность и реализация результатов работы. Работа носит теоретический характер. Все полученные в работе оценки продолжительности процедуры принятия решения являются неасимптотическими, а решающие правила явно заданными исходя из доступных априорных сведений о точности данных наблюдений, что позволяет их использовать на практике и строить статистические решения с близкими к оптимальным свойствами. Полученные оценки эффективности гарантийных статистических решений позволяют показать влияние априорных предположений о распределении наблюдений на вес наблюдений, которые рассматриваются как выбросы. Рассматриваемый подход позволяет проводить исследование областей изменения параметров статистической модели, в которых применение определенного статистического критерия является наиболее эффективным.

Апробация работы. Основные положения диссертационной работы докладывались и обсуждались на: 14-й Международной конференции по вероятности и статистике (Созопол, Болгария, 2008), XI Всероссийском симпозиуме по прикладной и промышленной математике (Кисловодск, 2010), 30-й – 33-й конференциях молодых учёных и специалистов ИППИ РАН: Информационные технологии и системы (2007)–(2010), научных семинарах ИППИ.

Публикации. По теме диссертации опубликовано 9 работ, в том числе 3 в изданиях из Списка ВАК.

Основные положения, выносимые на защиту:

1. Разработано понятие субоптимального статистического решающего правила, позволяющего получать близкие к оптимальным робастные статистические решения о проверке гипотез.

2. Субоптимальная последовательная процедура решения задачи проверки статистических гипотез с гарантийным решающим правилом при наличии априорной информации о равномерной относительной погрешности в плотности распределения наблюдений и ее многоэтапная модификация.

3. Свойства построенных субоптимальных процедур для задачи последо-

вательной проверки статистических гипотез с гарантийным решающим правилом при наличии априорной информации о скорости убывания хвостов распределений.

4. Установлены неасимптотические оценки функции риска и приведены примеры оценки эффективности статистических моделей для решения задачи проверки гипотез с учетом доступной априорной информации.

5. Анализ свойств χ^2–критерия как субоптимальной процедуры при наличии некоторого априорного свойства регулярности.

Объем работы. Диссертация состоит из введения, 4 глав заканчивающихся выводами по главе, заключения и списка использованной литературы. Работа изложена на 153 страницах машинописного текста, содержит 7 рисунков и 7 таблиц общим объемом 5 страниц. Список литературы включает 73 наименования.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении дана общая характеристика работы и приведено ее основное содержание.

В **первой главе** излагается история вопроса, формулируются основные определения и постановки задач.

Введем основные определения. Пусть $(X, \mathcal{B})$ измеримое пространство (для простоты предполагаем, что $X \subseteq \mathbb{R}$) с σ-аддитивной мерой μ; $x_1, \ldots, x_n, \ldots$ — наблюдения, т.е. значения случайных величин $x_i : \Omega \to X$, где $(\Omega, \mathcal{F}, \mathsf{P})$ — некоторое вероятностное пространство. В рамках настоящей диссертации считается, что x_i независимы, одинаково распределены и имеют плотность $f(x)$ относительно меры μ. Пусть на $(X, \mathcal{B})$ задан класс вероятностных мер $\mathcal{P} \ni \mathsf{P}_f$, абсолютно непрерывных относительно μ (соответствующее множество плотностей обозначим через $\mathcal{G}$), а также взаимно абсолютно непрерывных.

Если заданы функции $g_1, \ldots, g_m \in \mathcal{G}$, тогда задана задача проверки простых гипотез:

$$\mathcal{H}_1^0 : f = g_1, \ldots, \mathcal{H}_m^0 : f = g_m. \tag{1}$$

Для построения робастных правил проверки исходных простых гипотез необходимо задать окрестности распределений $g_1, \ldots, g_m$ исходя из априорных сведений о возможных ошибках. Эти окрестности $\mathcal{G}_1, \ldots, \mathcal{G}_m \subset \mathcal{G}$ должны удовлетворять условиям $\forall i \neq j,\ i, j = 1, \ldots, m,\ \mathcal{G}_i \cap \mathcal{G}_j = \varnothing,\ g_1 \in \mathcal{G}_1, \ldots, g_m \in \mathcal{G}_m$. Тогда на X задана задача проверки сложных гипотез, которая является расширением (1):

$$\mathcal{H}_1 : f \in \mathcal{G}_1, \ldots, \mathcal{H}_m : f \in \mathcal{G}_m. \tag{2}$$

В диссертации рассматриваются два вида множеств $\mathcal{G}_i$, которые называются окрестностями плотностей g_i.

В разделе 1.2 рассматривается ограниченное множество X, где в качестве множеств $\mathcal{G}_i$ рассматриваются множества функций $\widetilde{g}_i$, удовлетворяющих следующим условиям:

$$|\widetilde{g}_i(x) - g_i(x)| \leq \varepsilon g_i(x),\ x \in X, \tag{3}$$

$$\int\limits_X \widetilde{g}_i(x)\, d\mu(x) = 1 \tag{4}$$

В главе 1 показано, что при определенном значении параметров данные окрестности могут быть применены при наличии выбросов в модели Тьюки–Хубера:

$$\mathsf{P} = (1 - \lambda)\, \mathsf{P}_f + \lambda\, \mathsf{Q},$$

где $\lambda > 0$ — доля засоряющих предполагаемое теоретическое распределение P_f наблюдений, Q — распределение засоряющих наблюдений, которое считается неизвестным и может быть весьма произвольным; а также в ситуации, когда результат наблюдений искажается малым шумом ξ:

$$y = x + \xi,$$

где y — случайная величина, рассматриваемая как результат наблюдений, x — случайная величина, имеющая распределение P_f, а ξ — случайная величина, задающая шум.

В разделе 1.3 рассматривается неограниченное X, когда при формировании окрестностей g_i необходимо оценить сверху поведение плотности распределения на хвостах ($t_i^-(x)$ и $t_i^+(x)$), соответственно в качестве $\mathcal{G}_i$ рассматриваются множества функций $\widetilde{g}_i$, удовлетворяющих следующим условиям:

$$|\widetilde{g}_i(x) - g_i(x)| \leq \varepsilon g_i(x),\ x \in A_i = [a_i^-; a_i^+], \tag{5}$$

$$\widetilde{g}_i(x) \le t_i^-(x),\ x < a_i^-, \tag{6}$$

$$\widetilde{g}_i(x) \le t_i^+(x),\ x > a_i^+, \tag{7}$$

$$G_i := \int\limits_{a_i^-}^{a_i^+} g_i(x)\, d\mu(x) < 1. \tag{8}$$

$$\inf_{x\in A_i} g_i(x) \ge g_i^0 > 0. \tag{9}$$

Здесь отрезки A_i являются множествами, на которых сосредоточена бóльшая часть наблюдений, если справедлива гипотеза $\mathcal{H}_i$, а именно $G_i{=}\mathsf{P}_{g_i}(A_i)$.

В данной работе рассматриваются только последовательные процедуры с ограниченной величиной максимально возможной ошибки, т.е.

Определение 1.1. *Процедура $d = \langle\tau, \delta\rangle$ для задачи проверки гипотез* (2) *называется допустимой, если выполнены следующие условия:*

1. τ *— марковский момент остановки относительно естественно фильтрации $\Big\{\mathcal{F}_n = \sigma(x_1, \ldots, x_n)\Big\}$, т.е. $\forall\ n \in \mathbf{N}$ $\Big\{\omega : \tau(\omega) \le n\Big\} \in \mathcal{F}_n$ и $\mathsf{P}(\tau < \infty) = 1$.*

2. $\delta(\cdot)$ *является $\mathcal{F}_\tau$-измеримой случайной величиной.*

3. *Процедура d обеспечивает заданный уровень вероятности ошибки принятия неправильного решения, т.е.*

$$\forall\, i = 1, \ldots, m, \quad \sup_{g\in\mathcal{G}_i} \mathsf{P}_g(\delta(x_1, \ldots, x_\tau) \ne i) \le \alpha < 1.$$

Множество допустимых процедур для заданного уровня вероятности принятия ошибочного решения α обозначим через $\mathcal{D}(\alpha)$. В главе 1 показано, что построение допустимых процедур невозможно без наложения некоторых априорных предположений регулярности на множество мер $\mathcal{P}$.

Определение 1.2. *Функцией риска допустимой процедуры $d = \langle\tau, \delta\rangle$ для задачи проверки гипотез* (2) *является максимальная средняя продолжительность процедуры, т.е. при справедливости гипотезы $\mathcal{H}_i$:*

$$\mathsf{R}_i(d) = \sup_{g\in\mathcal{G}_i} \mathsf{E}_g\, \tau.$$

В разделе 1.6 вводятся многоэтапные процедуры принятия решений. Рассмотрение многоэтапных процедур обусловлено в практической точки зрения, поскольку часто (например, в медицинских исследованиях) невозможно выполнять наблюдения последовательно, или же выполнение наблюдений сериями существенно дешевле и быстрее. Пусть c — стоимость проведения одного наблюдения (если наблюдения выполняются сериями), а M — стоимость проведения этапа наблюдений.

Определение 1.4. *Процедура $d = \langle \tau, \delta \rangle \in \mathcal{D}(\alpha)$ для задачи проверки гипотез (2) называется допустимой многоэтапной процедурой, если выполнены следующие дополнительные условия:*
1. *задана последовательность продолжительности этапов наблюдений $N_1 > 0, N_2 > 0, \dots,$ а $\tau_0 = 0, \tau_k = \tau_{k-1} + N_k,\ k > 0, \dots$ — соответствующая последовательность моментов остановки этапов наблюдений; при этом N_k является $\mathcal{F}_{\tau_{k-1}}$-измеримой целочисленной случайной величиной;*
2. *момент завершения наблюдений τ совпадает с одним из моментов τ_k.*

Определение 1.5. *Функцией риска многоэтапной процедуры $d = \langle \tau, \delta \rangle$ для задачи проверки гипотез (2) назовем максимальную среднюю продолжительность процедуры с учетом стоимости этапов, т.е. при справедливости гипотезы $\mathcal{H}_i$*

$$\mathsf{R}_i(d) = \sup_{g \in \mathcal{G}_i} \mathsf{E}_g(c\tau + Mk^*),$$

где k^* — количество этапов.

В разделе 1.4 проводится обоснование постановки задачи, рассматриваемой в диссертации, на основе анализа свойств асимптотически оптимальных последовательных процедур. Это позволяет в разделе 1.5 сформулировать понятие субоптимальной процедуры.

На основании оценки снизу для средней продолжительности теста Вальда для любой допустимой процедуры $d \in \mathcal{D}(\alpha)$ решения задачи (2) справедлива следующая оценка снизу:

$$\mathsf{R}_i(d) \geq \frac{C(\alpha)}{\min\limits_{\{k:\, k \neq i\}} \mathsf{I}\,(g_i, g_k)},$$

где $\mathsf{I}(f, g)$ — информационное уклонение Кульбака–Лейбнера

$$\mathsf{I}(f, g) := \mathsf{E}_f z_{f,g}(x) = \int\limits_X z_{f,g}(x) f(x)\, d\mu(x),$$

где $z_{f,g}(x) = \ln \frac{f(x)}{g(x)}$, $C(\alpha) := -(1-\alpha)\ln\alpha - \alpha\ln(1-\alpha)$.

Главный член функции риска при $\alpha \to 0$ обозначим через

$$\mathsf{J}_i = \lim_{\alpha \to 0} \frac{R_i(d)}{|\ln \alpha|}.$$

Определение 1.3. *Назовем допустимую процедуру $d^* \in \mathcal{D}(\alpha)$ решения задачи проверки гипотез* (2) *субоптимальной, если*

$$\lim_{\varepsilon \to 0} \mathsf{J}_i(d^*) = \lim_{\varepsilon \to 0} \inf_{d \in \mathcal{D}(\alpha)} \mathsf{J}_i(d). \tag{10}$$

Смысл определения состоит в том, что при сжатии окрестностей заданных распределений g_i до нуля ($\varepsilon \to 0$) получаем асимптотически (при $\alpha \to 0$) оптимальную процедуру.

Во **второй главе** предложена процедура $d_0 =< \tau_0; \delta_0 >$ решения задачи (2), когда множества $\mathcal{G}_i$ заданы соотношениями (3) и (4), и исследованы свойства данной процедуры.

Пусть $g \in \mathcal{G} = \bigcup\limits_{i=1}^{m} \mathcal{G}_i$, тогда через $A(\mathsf{P})$ обозначим альтернативное множество, т.е. $A(g) := \bigcup\limits_{i \neq j} \mathcal{G}_i$, если $g \in \mathcal{G}_j$. Определим статистику $L_i(n)$ как

$$L_i(n) := \inf_{g \in A(g_i)} \sum_{k=1}^{n} z_{g_i, g}(x_k).$$

Для окрестностей (3) и (4) получаем, например при $m = 2$,

$$\begin{aligned} L_1(n) &= \sum_{i=1}^{n} \ln \frac{g_1(x_i)}{g_2(x_i)} - n \ln(1+\varepsilon), \\ L_2(n) &= -\sum_{i=1}^{n} \ln \frac{g_1(x_i)}{g_2(x_i)} - n \ln(1+\varepsilon), \end{aligned} \tag{11}$$

которые отличаются от классического теста Вальда только вычитанием из логарифма отношения правдоподобия каждого наблюдения постоянной величины, обусловленной неточностью задания статистической модели задачи.

Пусть $\beta = \frac{\alpha}{m-1}$, тогда момент остановки τ_0 определяется из условия

$$\tau_0 := \min \left\{ n : \max_{i=1,\dots,m} L_i(n) \geq -\ln \beta \right\} \tag{12}$$

и решающее правило: δ_0

$$\delta_0 = i, \quad \text{если} \quad L_i(\tau) \geq -\ln\beta. \tag{13}$$

Для процедуры d_0 получены следующие результаты.

Теорема 2.2. *Если существует константа $b > 0$, такая что*

$$\forall i, j = 1, \ldots, m \quad \mathsf{E}_{g_i} \left| \ln \frac{g_i(x)}{g_i(x)} \right|^{1+b} < \infty,$$

то процедура проверки гипотез (2) $d_0 \in D(\alpha)$, *если множества* $\mathcal{G}_i$ *задаются соотношениями* (3) *и* (4).

Для процедуры d_0 получена верхняя неасимптотическая граница для функции риска (для простоты обозначений формулировка приводится при $m = 2$).

Теорема 2.3. *Пусть для некоторого числа* $1 > b > 0$ *существует константа* C_1 *такая, что* $\mathsf{E}_{g_1}|\ln\frac{g_1(x)}{g_2(x)}|^{1+b} \leq C_1 < \infty$. *Тогда функция риска процедуры* d_0 *удовлетворяет условию*

$$\mathsf{R}_1(d_0) \leq \frac{|\ln\alpha| + K_1|\ln\alpha|^{1-b} + K_2|\ln\alpha|^{1-2b} + K_3}{(1-\varepsilon)\mathsf{E}_{g_1}(z_{g_1,g_2}(x))^+ - (1+\varepsilon)\mathsf{E}_{g_1}(z_{g_1,g_2}(x))^- - \ln(1+\varepsilon)},$$

где $a^+ := \max(a, 0), a^- := \max(-a, 0)$, *константы* K_1, K_2 *и* K_3 *задаются следующими формулами*

$$K_1 := \frac{(1+\varepsilon)}{b(1-b)((1-\varepsilon)\,\mathsf{E}_{g_1}(z_{g_1,g_2}(x))^+ - (1+\varepsilon)\,\mathsf{E}_{g_1}(z_{g_1,g_2}(x))^-)},$$

$$K_2 := \frac{(1+\varepsilon)(1-b)C_2}{b(1-2b)((1-\varepsilon)\,\mathsf{E}_{g_1}(z_{g_1,g_2}(x))^+ - (1+\varepsilon)\,\mathsf{E}_{g_1}(z_{g_1,g_2}(x))^-)},$$

$$K_3 := \frac{(1+\varepsilon)}{(1-\varepsilon)\,\mathsf{E}_{g_1}(z_{g_1,g_2}(x))^+ - (1+\varepsilon)\,\mathsf{E}_{g_1}(z_{g_1,g_2}(x))^-} \times$$
$$\times \left[\left(u_0 + \frac{1}{b u_0^b} \right) (u_0 + C_2 u_0^{1-b}) - \frac{u_0^{1-b}}{b(1-b)^2} - \frac{C_2 u_0^{1-2b}}{b(1-2b)} \right].$$

Если выполняется условие $\mathsf{E}_{g_1}|\ln\frac{g_1(x)}{g_2(x)}|^2 \leq C_1 < \infty$, *то существует константа* K_4, *такая что функция риска процедуры* d_0 *удовлетворяет условию*

$$\mathsf{R}_1(d_0) \leq \frac{|\ln\alpha| + K_4}{(1-\varepsilon)\,\mathsf{E}_{g_1}(z_{g_1,g_2}(x))^+ - (1+\varepsilon)\,\mathsf{E}_{g_1}(z_{g_1,g_2}(x))^- - \ln(1+\varepsilon)}. \tag{14}$$

Если выполняются условия

$$\inf_{x\in X} g_1(x) =: G_1^- > 0,\ \sup_{x\in X} g_1(x) =: G_1^+ < \infty$$

$$\inf_{x\in X} g_2(x) =: G_2^- > 0,\ \sup_{x\in X} g_2(x) =: G_2^+ < \infty,$$

то функция риска процедуры d_0 *удовлетворяет условию для задачи* (14) *с* $K_4 = \frac{G_1^+}{G_2^-}$.

Теорема 2.4. *Пусть для некоторого числа* $b > 0$ $\mathsf{E}_{g_i}|\ln\frac{g_1(x)}{g_2(x)}|^{1+b} \le C_i < \infty$. *Тогда процедура* d_0 *является субоптимальной, при этом скорость сходимости в* (10) *порядка* ε*:*

$$\mathsf{J}_i(d_0) \le \frac{1}{\mathsf{I}\,(g_i, g_j)} + \frac{1 + \mathsf{E}_{g_i}(z_{g_i,g_j}(x))^+ + \mathsf{E}_{g_i}(z_{g_i,g_j}(x))^-}{\mathsf{I}\,(g_i, g_j)^2}\,\varepsilon + o(\varepsilon),$$

$$\lim_{\varepsilon\to 0}\mathsf{J}_i(d_0) = \frac{1}{\mathsf{I}\,(g_i, g_j)} = \lim_{\varepsilon\to 0}\inf_{d\in\mathcal{D}(\alpha)}\mathsf{J}_i(d).$$

Во второй главе рассмотрена многоэтапная модификация процедуры d_0, которую обозначим $\widetilde{d}_0$. Предлагаемая процедура состоит из 3-х этапов, при этом может производится несколько итераций 3-го этапа; условие остановки $\widetilde{\tau}_0$ и решающее правило $\widetilde{\delta}_0$ совпадают, соответственно, с τ_0 и δ_0, определенными в (12), где n заменено на N_i, и (13). Для нее получена неасимптотическая верхняя граница для функции риска. Это позволяет обеспечить субоптимальность $\widetilde{d}_0$ при выполнении условия, что $M = O(|\ln\alpha|)$ (если $M = o(|\ln\alpha|)$, то ограничение на число этапов является слишком слабым ограничением при использовании понятия субоптимальности).

В **третьей глава** предложена процедура d_0 решения задачи (2), когда множества $\mathcal{G}_i$ заданы соотношениями (5)–(9), и исследованы свойства данной процедуры. В разделе 3.2 рассмотрен частный случай, когда плотности из $\mathcal{G}$ имеют экспоненциально убывающие хвосты, т.е. вместо (6) и (7) выполняются свойства

$$(1-\varepsilon)\, g_i(a_i^-)\, e^{r_i^-(x-a_i^-)} \le \widetilde{g}_i(x) \le (1+\varepsilon)\, g_i(a_i^-)\, e^{k_i^-(x-a_i^-)}, \qquad x < a_i^-, \quad (15)$$

$$(1-\varepsilon)\, g_i(a_i^+)\, e^{-r_i^+(x-a_i^+)} \le \widetilde{g}_i(x) \le (1+\varepsilon)\, g_i(a_i^+)\, e^{-k_i^+(x-a_i^+)}, \qquad x > a_i^+. \quad (16)$$

Для рассматриваемых множеств (15) и (16) статистика $L_i(n)$ строится следующим образом:

$$L_i(n) := \inf_{g \in A(g_i)} \sum_{i=1}^{n} z_{g_i^*, g}(x_i),$$

где

$$g_i^*(x) = \begin{cases} g_i(a_i^-)(1+c_i)\, e^{k_i^-(x-a_i^-)}, & \text{если } x < a_i^-; \\ g_i(x)(1+c_i), & \text{если } a_i^- \le x \le a_i^+; \\ g_i(a_i^+)(1+c_i)\, e^{-k_i^+(x-a_i^+)}, & \text{если } x > a_i^+. \end{cases} \tag{17}$$

Значение c_i определяется из условия нормировки $\int\limits_{-\infty}^{+\infty} g_i^*(x) d\mu(x) = 1$. При этом должно выполняться условие $|c_i| \le \varepsilon$, которое следует из (15) и (16). Момент остановки τ_0 и δ_0 определяются аналогично (12) и (13).

В диссертации приведены формулы для расчета статистик $L_i(n)$ для множеств $\mathcal{G}_i$, задаваемых (5), (15), (16), (8) и (9). Для экономии места приведем один пример формулы для статистики $L_1(n)$:

$$\begin{aligned} L_1(n) = n \ln\left(\frac{1+c_1}{1+\varepsilon}\right) + \sum_{i: x_i \le a_1^-} \left(\ln\left(\frac{g_1(a_1^-)}{g_2(a_2^-)}\right) + k_1^-(x_i - a_1^-) - k_2^-(x_i - a_2^-) \right) + \\ + \sum_{i: a_1^- < x_i \le a_2^-} \left(\ln\left(\frac{g_1(x_i)}{g_2(a_2^-)}\right) - k_2^-(x_i - a_2^-) \right) + \sum_{i: a_2^- < x_i \le a_1^+} \ln\left(\frac{g_1(x_i)}{g_2(x_i)}\right) + \\ + \sum_{i: a_1^+ < x_i \le a_2^+} \left(\ln\left(\frac{g_1(a_1^+)}{g_2(x_i)}\right) - k_1^+(x_i - a_1^+) \right) + \\ + \sum_{i: x_i > a_2^+} \left(\ln\left(\frac{g_1(a_1^+)}{g_2(a_2^+)}\right) - k_1^+(x_i - a_1^+) + k_2^+(x_i - a_2^+) \right). \end{aligned}$$

Таким образом, вид статистик $L_i(n)$ существенно усложнился по сравнению с (11). Новая статистика зависит от границ скорости убывания хвостов распределений.

Рассмотрим некоторые примеры соотношения между параметрами модели, которые оказывают качественное влияние на вид статистик $L_1(n)$ и $L_2(n)$. Первый случай — совпадение минимальных скоростей убывания хвостов распределений для обеих гипотез. В этом случае приращения величин $L_i(n)$ ограничены. Таким образом мы получаем вывод, который часто под-

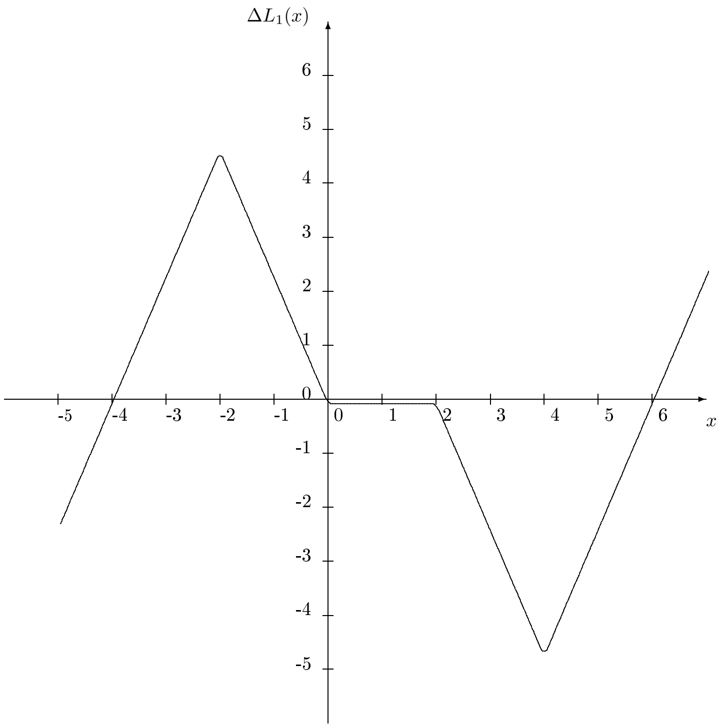

Рис.3.4. График $\Delta L_1(x)$

тверждается на практике: выбросы в наблюдениях нужно учитывать с ограниченными весами. Второй случай — когда на $-\infty$ минимальная скорость убывания хвостов распределений у первой гипотезы больше, чем у второй, а на $+\infty$ наоборот. Вид приращения величины $L_1(n)$ приведен на рис. 3.4. В этом случае мы получаем совершенно другой вывод: выбросы в наблюдениях нельзя отбрасывать, более того, вес таких наблюдений будет отрицательным и может принимать неограниченные значения.

Приведенными примерами не описывается все множество вариантов соотношений между параметрами рассматриваемой модели, которые приводят к различному виду статистик $L_i(n)$. Однако приведенные примеры указывают на принципиальные особенности статистик $L_i(n)$ при наличии экспоненциальной скорости убывания хвостов распределений: при определенных ситуациях можно использовать ограниченные функции для построения приращений статистик $L_i(n)$, обеспечивающих робастное решение задачи проверки гипотез, а в других случаях использование ограниченных функций для статистик $L_i(n)$, обеспечивающих робастное решение задачи проверки гипотез, невозможно.

Введена характеристика $I_1(p)$, позволяющая получить аналог утверждения (14). Кроме того, доказано, что в рассматриваемом случае d_0 является субоптимальной процедурой. Вместе с тем, этот результат принципиально отличается от результатов главы 2. В предыдущем случае границы субоптимальной процедуры и процедуры Вальда проверки гипотез (1) асимптотически совпадали. Сейчас это может не выполняться, если у наблюдаемого распределения скорость убывания хвоста существенно отличается от соответствующей характеристики в (17).

В разделе 3.3 рассмотрен общий случай, когда окрестности задаются условиями (5)–(9). Основной результат состоит в следующем.

Теорема 3.4. *Пусть для $p_1 \in \mathcal{G}_1$ для некоторого числа $1 > b > 0$ существует константа C_1 такая, что $\mathsf{E}_{p_1}|\Delta L_1|^{1+b} \leq C_1 < \infty$. Тогда существуют константы K_1, K_2 и K_3, такие что функция риска процедуры d_0 удовлетворяет условию*

$$\mathsf{R}_1(d_0) \leq \frac{|\ln\alpha| + K_1|\ln\alpha|^{1-b} + K_2|\ln\alpha|^{1-2b} + K_3}{I_1(p_1)}.$$

Если для $p_1 \in \mathcal{G}_1$ существует константа C_1 такая, что выполняется условие $\mathsf{E}_{p_1}|\Delta L_1|^2 \leq C_1 < \infty$, то существует константа K_4, такая что функция риска процедуры d_0 удовлетворяет условию

$$\mathsf{R}_1(d_0) \leq \frac{|\ln\alpha| + K_4}{I_1(p_1)}. \tag{18}$$

Если выполняются условия $\sup_{x\in X} \Delta L_1 \leq K_5$, то функция риска процедуры d_0 удовлетворяет условию (18) с $K_4 = K_5$.

Аналогичное утверждение справедливо и для $p_2 \in \mathcal{G}_2$.

В **четвертой главе** диссертации приведены результаты численных исследований эффективности предлагаемых процедур. Проведен сравнительный анализ эффективности оптимальной последовательной процедуры Вальда проверки простых гипотез (2) и субоптимальных процедур проверки соответствующих сложных гипотез при различных предположениях об априорных сведениях о возможных погрешностях в распределении наблюдений. Кроме сравнительного анализа продолжительности процедур вычисляется и точность принимаемого решения. Рассчитанное среднее значение продолжительности обозначается через R для субоптимальной процедуры и R_W для процедуры Вальда. Вероятность принятия неверной гипотезы $\mathcal{H}_2$ (при моделировании предполагаем, что $m = 2$ и справедлива гипотеза $\mathcal{H}_1$) обозначается через p для субоптимальной процедуры и p_W для процедуры Вальда. В качестве иллюстрации приводятся результаты моделирования (10000 экспериментов) субоптимальной процедуры, когда окрестности исходных распределений g_i задаются соотношениями (3) и (4).

Полученные результаты показывают, что применение классической процедуры Вальда может приводить к тому, что она не обеспечивает заданные уровни ошибок. В таблице выделены те результаты, когда вероятность отклонения гипотезы $\mathcal{H}_1$ оказалась выше заданного уровня α. Приведенные результаты не могут быть объяснены статистической погрешностью, поскольку значение α не попадает в доверительный интервал для P_W уровня 0.995.

Таблица 4.1 *Результаты численного исследования свойств субоптимальности при равномерной оценке плотности распределений*

α	ε	R_W	R	p_W	p
0.01	0.01	10.21	10.61	0.0086	0.0085
0.01	0.05	11.06	12.40	**0.0130**	0.0074
0.001	0.05	16.59	18.39	**0.0021**	0.0009
0.01	0.1	12.16	16.35	**0.0201**	0.0051
0.001	0.1	18.36	23.79	**0.0041**	0.0005
0.01	0.15	13.37	23.63	**0.0296**	0.0030
0.001	0.15	20.64	33.82	**0.0053**	0.0005

В разделе 4.6 приведен пример упрощения модели, когда редукция приводит к увеличению информационного расстояния между гипотезами.

В разделе 4.7 рассматривается связь между χ^2-критерием и субоптимальными решающими правилами, рассматриваемыми в работе. Показано, что при наличии дополнительного условия регулярности: монотонности отношения правдоподобия, статистика χ^2-критерия в определенном смысле близка к статистике $L_2(n)$, ориентированной на принятие альтернативы, а статистика критерия знаков — к статистике $L_1(n)$, ориентированной на принятие основной гипотезы. Также получены поправочные коэффициенты для статистики χ^2-критерия, позволяющие обеспечить его робастность в предположениях главы 2.

В **заключении** сформулированы основные результаты диссертации:

1. Построены статистические модели окрестностей распределений, учитывающие априорную информацию о возможных ошибках в результатах измерений и о неопределенностях в описании распределения результатов наблюдений. Показано, что при этом возникают непараметрические множества вероятностных распределений, поэтому робастные статистические решения о справедливости одной из простых гипотез приводят к задаче проверки непараметрических гипотез. Показано, что в случае ограниченного множества возможных значений X возможны статистические модели окрестностей распределений с равномерной относительной погрешностью определения плотности распределений.

2. Введено понятия субоптимального статистического решающего правила, позволяющего выделять из всего множества допустимых решений близкие к оптимальным робастные статистические решения о проверке гипотез.
3. Построены субоптимальные решающие правила в задаче последовательной проверки статистических гипотез с гарантийным решающим правилом при наличии априорной информации об равномерной относительной погрешности в плотности распределения наблюдений, включая многоэтапные процедуры, а так же при наличии априорной информации об экспоненциальной скорости убывания хвостов распределений.
4. Для построенных процедур получены неасимптотические оценки функции риска. Это позволяет использовать полученные оценки для выбора статистической модели, обеспечивающей наиболее эффективное решение задачи проверки гипотез с учетом доступной априорной информации.
5. Проведено исследование влияния априорной информации о скорости убывания хвостов распределений на гарантийное решающее правило. Показано, что традиционные методы обработки результатов наблюдений с целью обеспечения робастности последующих статистических решений при определенных условиях могут быть неэффективными и не обеспечивать робастность этих решений.
6. Проведенное численное моделирование подтвердило теоретические положения об обеспечении предлагаемыми статистическими решениями заданного уровня ошибок. Субоптимальные статистические решения при малых погрешностях в значениях параметров модели обеспечивают практически те же или близкие значения функции риска, что и стандартный тест Вальда.
7. Показано, что при наличии условия монотонности отношения правдоподобия, выполняющегося с определенной точностью, критерий χ^2 построен на основании статистики, которая асимптотически является субоптимальной, что объясняет часто наблюдаемую эффективность этого критерия. Введение поправочных коэффициентов, вытекающих из соответствующей задачи построения субоптимальных решающих правил, позволяет устранить типичные погрешности критерия: низкую эффективность для малых и очень больших выборок. Кроме того, подход на основании построения субоптимальных процедур позволяет получить обоснование правил группирования данных наблюдений.

Работы автора по тебе диссертации

1. Цитович Ф.И. Субоптимальные последовательные правила проверки гипотез при слабых возмущениях // Обозрение прикладной и промышленной математики. 2008. Т.15, № 6. С. 1141.
2. Цитович Ф.И. Некоторые субоптимальные последовательные правила проверки гитопез // Сборник трудов 30-й конференции молодых учёных и специалистов ИППИ РАН: Информационные технологии и системы ИТиС'07. М.: ИППИ, 2007. С. 110–115.
3. F. Tsitovich Supoptimal Multistage Nonparametric Hypotheses Test // Pliska Stud. Math. Bulgaria. 2009. Vol. 19, pp. 269–282.
4. Цитович Ф.И. Многоэтапные процедуры проверки статистических гипотез и субоптимальнoть // Сборник трудов 31-й конференции молодых учёных и специалистов ИППИ РАН: Информационные технологии и системы ИТиС'08. М.: ИППИ. 2008. С. 386–390.
5. Цитович Ф.И. Субоптимальные последовательные правила проверки непараметрических гипотез о распределениях с экспоненциально убывающими хвостами // Сборник трудов 32-й конференции молодых учёных и специалистов ИППИ РАН: Информационные технологии и системы ИТиС'09. М.: ИППИ. 2009. С. 416–422.
6. Цитович Ф.И. Многоэтапные процедуры проверки статистических гипотез и субоптимальнoть // Сборник трудов 33-й конференции молодых учёных и специалистов ИППИ РАН: Информационные технологии и системы ИТиС'10. М.: ИППИ. 2010. С. 252–257.
7. Цитович Ф.И. Субоптимальные последовательные правила проверки гипотез при слабых возмущениях // Обозрение прикладной и промышленной математики. 2010. Т.17, № 2. С. 315—316.
8. Цитович Ф.И. Робастность и субоптимальные статистические решения в задачах последовательной проверки гипотез // Обозрение прикладной и промышленной математики. 2010. Т.17, № 4. С. 603—604.
9. Цитович Ф.И. Свойства субоптимальных последовательных правил проверки непараметрических гипотез о распределениях с экспоненциально убывающими хвостами // Информационные процессы. Т. 10, № 2. 2010. С. 181-196.