САНКТ-ПЕТЕРБУРГСКОЕ ОТДЕЛЕНИЕ МАТЕМАТИЧЕСКОГО
ИНСТИТУТА им. В. А. СТЕКЛОВА РОССИЙСКОЙ АКАДЕМИИ НАУК

На правах рукописи

Симушкин Дмитрий Сергеевич

Статистические критерии с ограничениями на d-риски

Специальность: 01.01.05 – теория вероятностей и математическая статистика

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата физико-математических наук

Санкт-Петербург – 2020

Работа выполнена на кафедре математической статистики Казанского (Приволжского) Федерального Университета.

Научный руководитель:	**Володин Игорь Николаевич** доктор физико-математических наук, профессор кафедры математической статистики КФУ
Официальные оппоненты:	**Бернштейн Александр Владимирович** доктор физико-математических наук, профессор, «Сколковский институт науки и технологии»
	Малов Сергей Васильевичч кандидат физико-математических наук, ведущий научный сотрудник лаборатории «Центр генной биоинформатики им Ф. Добржанского», ФГБОУ ВО «Санкт-Петербургский государственный университет»
Ведущая организация:	Московский государственный университет им. М.В. Ломоносова

Защита состоится «___» _____ 2020 г. в ____ часов на заседании диссертационного совета Д 002.202.01 в ФГБУН «Санкт-Петербургское отделение Математического института им. В. А. Стеклова Российской академии наук» по адресу: 191023, Санкт-Петербург, наб. р. Фонтанки, д. 27.

С диссертацией можно ознакомиться в библиотеке ФГБУН «Санкт-Петербургское отделение Математического института им. В. А. Стеклова Российской академии наук» или на сайте по адресу http://pdmi.ras.ru/

Автореферат разослан «___» _____ 2020 г.

Учёный секретарь
диссертационного совета Д 002.202.01,
доктор физико-математических наук, А. Ю. Зайцев

Общая характеристика работы

Актуальность темы исследования диссертации. В конце XX века, с подачи Л.Н. Большева, усилиями И.Н. Володина и его учеников [1, 2, 3] получил развитие так называемый d-апостериорный подход к проблеме гарантийности статистического вывода. В этом подходе риск любого статистического правила (d-риск) вычисляется как условное среднее возможных потерь среди экспериментов закончившихся принятием одного и того же решения. Из такого определения риска следует, что d-апостериорный подход применим только к ситуациям, когда имеется реальная последовательность статистических экспериментов, в каждом из которых необходимо принять решение об изучаемом объекте. Зачастую в таких ситуациях можно предположить, что характеристика объекта, относительно которой принимается решение, изменяется случайно от эксперимента к эксперименту в соответствии с некоторым априорным распределением.

В [1] была предложена универсальная последовательная d-гарантийная процедура, которую применительно к задаче различения двух сложных гипотез $\mathrm{H}_0 : \theta \in \Theta_0$, $\mathrm{H}_1 : \theta \in \Theta_1$ относительно неизвестного параметра распределения наблюдений $X^{(n)} = (X_1, X_2, \ldots)$ можно описать следующим образом. Универсальная процедура останавливается на шаге с номером n после получения выборочного вектора $x^{(n)} = (x_1, \ldots, x_n)$, если апостериорная вероятность какой-либо из гипотез $\mathbf{P}\big(\vartheta \in \Theta_j \,|\, x^{(n)}\big) \geqslant 1 - \beta_j$ (β_0, β_1 — заданные ограничения на d-риск). Вопрос о замкнутости момента прекращения наблюдений и о конечности его математического ожидания до сих пор не был систематически изучен.

Другая последовательная процедура была рассмотрена в статье [4]. Остановка этой процедуры происходит в тот момент, когда впервые значение статистики вклада выйдет за (постоянные) границы, приведённые в этой статье. Были установлены оптимальные (в асимптотическом смысле при $\beta_0, \beta_1 \to 0$) свойства этой процедуры. Однако вопрос сравнения этой процедуры с универсальной d-гарантийной процедурой также оставался открытым.

В работе [5] предложены способы приближённого вычисления необходимого объёма выборки (НОВ) при d-гарантийном различении двух односторонних гипотез для двух схем асимптотического анализа. Абсолютная точность этих асимптотик при малых значениях НОВ не велика. Уточнение этих при-

[1] Володин И. Н., Симушкин С. В. О d-апостериорном подходе к проблеме статистического вывода// 3-я Виль.Конф.Теор.Вер. и Мат.Стат. – 1981, Vol. **1**. – с. 100–101.

[2] Володин И. Н. Гарантийные процедуры статистического вывода (определение объёма выборки)// *Иссл.Прикл.Матем.Информ.* – 1984, № 10. – с. 13–53.

[3] Володин И. Н., Новиков Ан. А., Симушкин С. В. Гарантийный контроль качества: апостериорный подход// *Обозрение Прикладной и Промышленной Математики* – 1994, т. **1**, № 2. – с. 1–32.

[4] Новиков Ан. А. Асимптотическая оптимальность последовательного d-гарантийного критерия// *Теор.Вер. и Примен.* – 1987, т. **32**, № 2, с. 387–391.

[5] Volodin I. N., Novikov An. A. Asymptotics of the necessary sample size in testing parametric hypothesis: d-posterior appropach// *Mathematical Methods of Statistics* – 1998, Vol. **7**, № 1, – p. 111–121.

ближённых формул в общем случае весьма затруднительно, поскольку требует асимптотических разложений апостериорного распределения с равномерными по параметру остатками.

В последнее время большую популярность приобрели методы статистического анализа в ситуациях так называемого множественного тестирования, в которых гарантийность процедур связывается с относительной частотой ложных «открытий» (отвержений нулевой гипотезы). Предложенный в [6] алгоритм позволяет контролировать среднюю долю этой частоты — так называемый показатель FDR (false discovery rate). В байесовской постановке аналогичную характеристику (pFDR) предложил Storey J. D.[7] В частности его методология была реализована в [8] для задачи сравнения генов в различных группах пациентов. Следует сказать, что в байесовской постановке показатель pFDR есть не что иное, как d-риск первого рода в d-апостериорном подходе. Таким образом, открывается широкая область применения методов d-апостериорного подхода на практике. Преимущество d-апостериорного подхода заключается в возможности осуществления контроля не только за относительной частотой ложных открытий, но также и за относительной частотой ложных принятий нулевой гипотезы и, кроме того, даёт возможность построения процедур различения более двух гипотез.

Целью исследования диссертационной работы является анализ свойств d-гарантийных процедур различения двух односторонних гипотез $\mathrm{H}_0 : \theta \in \Theta_0 = (-\infty, \theta_0\,]$ и $\mathrm{H}_1 : \theta \in \Theta_1 = (\theta_0, \infty)$ о действительном параметре θ в рамках d-апостериорного подхода. А именно: исследование свойств момента остановки универсальной процедуры (замкнутость момента остановки и конечность его математического ожидания) и сравнение этой процедуры с последовательной процедурой на статистике вклада и d-гарантийной процедурой с фиксированным числом наблюдений; уточнение асимптотических формул необходимого объёма выборки; разработка методики применения d-апостериорного подхода к задачам множественного тестирования.

Методы исследования. При исследовании свойств процедур использовались классические методы математического анализа и теории вероятностей, в частности, закон повторного логарифма, тождество Вальда, разложение функции распределения в ряд Эджворта, асимптотические представления для функций распределения, разложение в ряд Тейлора. Сравнение последовательных процедур производилось методом стохастического моделирования.

[6] Benjamini Y., Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing// *Journal of the Royal Statistical Society: Series B.* – 1995, Vol. **57**, № 1. – p. 289–300.

[7] Storey J. D. A direct approach to false discovery rates// *Journal of the Royal Statistical Society: Series B* – 2002, Vol. **64**, № 1. – p. 479–498

[8] Efron B. *Large-Scale Inference. Empirical Bayes methods for estimation, testing, and prediction*/ B. Efron. – Cambridge, New York: Cambridge University Press, 2010. – 321 p.

Научная новизна диссертационной работы.

Показано, что универсальная последовательная d-гарантийная процедура есть процедура вальдовского типа относительно условного правдоподобия.

Доказана замкнутость момента остановки универсальной последовательной процедуры для вероятностных моделей, для которых область продолжения наблюдений имеет границы параболического типа. Показано, что границы подобного типа возникают для рассмотренных в диссертации вероятностных моделей.

Установлена конечность условного математического ожидания момента остановки универсальной последовательной процедуры для нормально-нормальной модели при истинном значении параметра, отличном от границы различаемых гипотез.

Доказано, что при истинном значении параметра, равном границе различаемых гипотез, среднее значение момента остановки универсальной последовательной процедуры равно бесконечности.

Предложен вариант универсальной последовательной d-гарантийной процедуры с усечённым моментом остановки и численно проведено сравнение этой процедуры с существующими последовательными процедурами, разработанными в рамках d-апостериорного подхода.

Найдены уточнения асимптотических формул необходимого объёма выборки для двух вероятностных моделей в двух асимптотических схемах, существенным образом снижающие величину ошибки приближённых формул.

Предложено уточнение вероятностной модели Эфрона в задаче множественного тестирования, для которой найден вид оптимальной d-гарантийной процедуры. Дана общая схема построения процедур множественного сравнения.

Основные положения выносимые на защиту

1. Момент остановки универсальной последовательной процедуры для вероятностных моделей, для которых область продолжения наблюдений имеет границы параболического типа, замкнут. В частности, таковыми являются нормально-нормальная, бета-Бернулли и гамма-показательная модели.

2. Условное математическое ожидание момента остановки универсальной последовательной процедуры для нормально-нормальной модели конечно при истинном значении параметра, отличном от границы различаемых гипотез, и бесконечно в противном случае.

3. Найдены уточнения асимптотических формул необходимого объёма выборки в схеме стягивающегося априори и в схеме с жёсткими ограничениями для нормально-нормальной, бета-Бернулли и гамма-показательной вероятностных моделей.

4. Методика применения d-апостериорного подхода к задачам множественного тестирования позволяет контроллировать на заданном уровне вероятности d-апостериорных ошибок первого и второго рода, а также даёт возможность по-

строения d-гарантийных процедур различения многих гипотез. Кроме того, эта методика позволяет управлять наблюдениями для достижения заданных уровней на вероятности ошибок.

Практическая значимость работы. Работа, в основном, носит теоретический характер. Часть результатов может быть полезна при организации статистического контроля качества массовой продукции, решении задач медицинской диагностики и задач множественного тестирования.

Апробация результатов исследования. Основные результаты диссертации представлялись на: международной конференции «Systems Biology and Medicine», SysPatho Workshop, St. Petersburg, 2012, 11-й международной Вильнюсской конференции, Вильнюс, 2014, XX-й Всероссийской Школе-коллоквиуме по стохастическим методам, Йошкар-Ола, 2013, Международной конференции по алгебре, анализу и геометрии и их приложениям, Казань, 2016, Международной конференции по теории вероятностей и математической статистике, Казань, 2017, городском семинаре г. Санкт-Петербурга по теории вероятностей и математической статистике, научном семинаре кафедры математической статистики ВМК МГУ, ежегодных конференциях К(П)ФУ.

Личный вклад автора. Диссертантом совместно с научным руководителем проводились выбор темы, планирование работы, постановка задачи и обсуждение полученных результатов. Результаты первой главы получены автором диссертации. Результаты второй главы получены совместно с научным руководителем Володиным И.Н. и Симушкиным С.В.

Публикации. По теме диссертации опубликовано 8 печатных работ, 4 — в изданиях из перечня рецензируемых научных журналов ВАК, 3 — в международных изданиях, индексируемых в базе данных Scopus и WoS, 1 работа принята к рассмотрению, 4 — тезисы международных конференций.

Объём и структура диссертации

Диссертационная работа состоит из введения, двух глав, заключения, списка обозначений, списка таблиц и списка литературы. Материал изложен на 135 страницах, включает 17 таблиц, 5 рисунков. Список использованных литературных источников содержит 70 наименований.

Содержание работы.

В разделе **1.1 главы 1** описываются основные положения теории принятия решений с ограничениями на d-апостериорные вероятности ошибок. Пусть требуется проверить гипотезу $\mathrm{H}_0 : \theta \in \Theta_0$ о параметре θ, индексирующем распределение $\mathbf{P}_\theta$ наблюдаемой случайной величины X. Пусть это распреде-

ление описывается плотностью $f(\cdot\,|\,\theta)$ относительно некоторой сигма-конечной меры μ. Предположим, что значение θ есть реализация случайной величины ϑ с некоторой функцией распределения G (плотностью g). Если решение d_0 в пользу H_0 или решение d_1 в пользу альтернативы $\mathrm{H}_1 : \theta \notin \Theta_0$ принимается посредством решающей функции δ на основе последовательности наблюдений $X^{(\nu)} = (X_1, \ldots, X_\nu)$ с моментом остановки ν, то d-риск 1-го рода δ определяется как условная вероятность

$$\mathcal{R}_1(\delta) = \mathbf{P}\big(\vartheta \in \Theta_0 \,|\, \delta = d_1\big),$$

где $\mathbf{P}$ — совместное распределение наблюдений и неизвестного параметра ϑ. Аналогично, d-риск 2-го рода $\mathcal{R}_0(\delta) = \mathbf{P}\big(\vartheta \notin \Theta_0 \,|\, \delta = d_0\big)$. Устанавливается лемма, которая существенно используется в дальнейших построениях.

Лемма 1.1. *Пусть δ — некоторая решающая функция в задаче различения двух гипотез $H_0 : \theta \in \Theta_0$, $H_1 : \theta \notin \Theta_0$, $\Pi_0 = \mathbf{P}\big(\vartheta \in \Theta_0\big)$ — априорная вероятность Θ_0, $\Psi(d_0) = \mathbf{P}\big(\delta = d_0\big)$ — безусловная вероятность принятия решения d_0. Тогда d-риски решающей функции δ связаны равенством*

$$\mathcal{R}_1(\delta) = 1 - \frac{1 - \Pi_0 - \mathcal{R}_0(\delta)\Psi(d_0)}{1 - \Psi(d_0)}\,.$$

В разделе **1.2** ставится задача построения критерия, основанного на фиксированном числе наблюдений ($\nu \equiv n$), с необходимым объёмом выборки (НОВ) $n = n^*$, при котором этот критерий гарантирует заданные ограничения на обе d-апостериорные вероятности ошибок:

$$\mathcal{R}_0(\delta) \leqslant \beta_0, \quad \mathcal{R}_1(\delta) \leqslant \beta_1.$$

Рассматривается задача различения гипотез $\mathrm{H}_0 : \theta \leqslant \theta_0$ и $\mathrm{H}_1 : \theta > \theta_0$ в рамках трёх популярных вероятностных моделей: а) модель N–N с нормальным (θ, σ^2) распределением наблюдений и нормальным распределением выводного параметра θ, б) модель G–E с показательным распределением наблюдений (неизвестный параметр θ — параметр интенсивности) и гамма-распределением θ, в) модель B–B с бернуллиевским распределением наблюдений (θ — вероятность «успеха») и бета-распределением θ.

Теорема 1.1. *Для задачи различения гипотез $H_0 : \theta \leqslant \mu$ и $H_1 : \theta > \mu$ в рамках модели* N–N *, где μ — априорное среднее, с одинаковыми ограничениями $\beta_0 = \beta_1 = \beta$ на d-риски, необходимый объём выборки*

$$n^* = \left\lceil \frac{\sigma^2}{\tau^2 \operatorname{tg}^2(\pi\beta)} \right\rceil,$$

где $\lceil a \rceil$ — целая часть числа a с округлением вверх.

Доказательство этого утверждения существенно опирается на утверждение **леммы 1.3**, в которой задача отыскания НОВ сводится к решению уравнения интегрального типа.

Теорема 1.2. *В нормально-нормальной модели с априорной дисперсией* $\tau^2 \to \infty$ *при различении гипотез* $H_0 : \theta \leqslant \theta_0$ *и* $H_1 : \theta > \theta_0$ *необходимый объём выборки* $n^* \to 1$.

В этом утверждении нормальное распределение с бесконечной дисперсией выступает в роли «равномерного» априорного распределения, выбираемого обычно как наименее информативное распределение параметра.

Переходя к модели G–E, сначала (**лемма 1.4**) приводится вид апостериорного распределения параметра ϑ при фиксированном значении выборочной суммы $S_n = \sum_1^n X_i$, а также безусловное распределение S_n представляется через функцию распределения бета-закона.

Построение последовательных и асимптотических процедур для модели G–E основывается на асимптотическом представлении для обратной функции гамма-распределения с параметром формы, стремящимся к бесконечности.

Лемма 1.5. *Пусть* $t_\gamma = \Phi^{-1}(\gamma)$ *— квантиль порядка* γ *стандартного нормального* $(0,1)$ *распределения. Тогда обратная функция гамма-закона* $\mathscr{G}(n,1)$ *имеет асимптотическое (при* $n \to \infty$ *) представление*

$$\mathbb{G}^{-1}(\gamma\,;\,n,1) = n + \sqrt{n}t_\gamma + \frac{1}{3}\big(t_\gamma^2 - 1\big) + o(1).$$

В разделе **1.2.1** изучается асимптотика НОВ n^* при различении гипотез $\mathrm{H}_0 : \theta \leqslant \theta_0$, $\mathrm{H}_1 : \theta > \theta_0$ в ситуации, когда ограничения $\beta_0, \beta_1 \to 0$. В **леммах 1.6** и **1.7** даются упрощённые представления для асимптотических формул из[5]. Для моделей N–N и G–E получены уточнения этой асимптотики.

Введём следующие обозначения: Φ, ϕ — функция распределения и, соответственно, функция плотности стандартного нормального $(0,1)$ закона, $\rho = \beta_1/\beta_0$, $W(c) = \phi(c) + c\Phi(c)$, $c \in (-\infty, \infty)$, $\Pi_0 = \Phi\big((\theta_0 - \mu)/\tau\big)$ — априорная вероятность справедливости нулевой гипотезы, $g_0 = \tau^{-1}\phi((\theta_0 - \mu)/\tau)$ — значение априорной плотности в граничной точке. Выберем c_0 из уравнения $W(c_0)\big(\Pi_0(1+\rho) - \rho\big) = c_0\Pi_0$ и положим $\Delta_0 = \rho - (1+\rho)\Pi_0$, $Z = \Pi_0 + \Delta_0\Phi(c_0)$.

Теорема 1.3. *Пусть* $\beta_0, \beta_1 \to 0$ *так, что* $\beta_1/\beta_0 = \rho > 0$. *Тогда в рамках модели N–N необходимый объём выборки*

$$n^* = \left\lceil \sigma^2 \left(\frac{W(c_0) g_0}{\Pi_0} \, \frac{1}{\beta_0} + V(\beta_0) \right)^2 \right\rceil,$$

где $\lceil a \rceil$ *— наименьшее целое число, не меньше* a, *и*

$$\lim_{\beta_0 \to 0} V(\beta_0) = \frac{Q}{Z}\,,$$

$$Q = \frac{\mu\Delta_0}{2\tau^2}\big(2c_0\Phi(c_0) + \phi(c_0)\big) + c_0\left(\frac{\mu\Pi_0}{\tau^2} - g_0 + (1+\rho)g_0\Phi(c_0)\right).$$

Для модели G–E справедливо аналогичное утверждение. Пусть априорная плотность $g(\theta\,;\,\lambda, a)$ есть гамма-плотность с параметром формы λ и параметром интенсивности a, Π_0 — априорная вероятность справедливости нулевой гипотезы, $g_0 = g(\theta_0\,;\,\lambda, a)$ — значение априорной плотности в граничной точке. Определим c_0, Δ_0 и Z как и выше.

Теорема 1.4. *Для гамма-показательной модели при различении гипотез* $H_0 : \theta \leqslant \theta_0$, $H_1 : \theta > \theta_0$ *и ограничениях* $\beta_1 = \rho\beta_0 \to 0$ *НОВ*

$$n^* = \left\lceil \frac{1}{\theta_0^2}\left(\frac{W(c_0)g_0}{\Pi_0}\,\frac{1}{\beta_0} + V_1(\beta_0)\right)^2 \right\rceil,$$

где

$$\lim_{\beta_0\to 0} V_1(\beta_0) = \frac{Q_1}{Z},$$

$$\begin{aligned} Q_1 = \; & -\frac{\Delta_0}{6}(3a\theta_0 + 1 - 3\lambda)\phi(c_0) + \rho\big(\theta_0 g_0 - (a\theta_0 - \lambda)(1-\Pi_0)\big)c_0 - \\ & - \theta_0 g_0(1+\rho)\Phi(-c_0)c_0 \; + \Delta_0(a\theta_0 - \lambda)\Phi(-c_0)c_0\,. \end{aligned}$$

В разделе **1.2.2** изучаются возможности применения к описанным трём вероятностным моделям приближённой формулы для НОВ n^*, полученной в статье [9] в схеме стягивающегося априори. В этой схеме предполагается, что априорная плотность может быть представлена в виде $g(\theta) = \frac{1}{\tau}\widetilde{g}\big((\theta-\theta_0)/\tau\,;\,\tau\big)$, где функция $\widetilde{g}(\cdot\,;\,\tau) \to \widetilde{g}(\cdot\,;\,0)$ при $\tau \to 0$.

В лемме **1.8** устанавливается, что для модели N–N асимптотические формулы [9] дают способ вычисления точного значения НОВ n^*. Для вероятностных моделей G–E и B–B функция $\widetilde{g}$ может быть выбрана как плотность нормального распределения с единичной дисперсией и математическим ожиданием, зависящим от величины отклонения априорного среднего от граничной точки θ_0 (леммы **1.9**, **1.10**). Численные расчёты показали, что без учёта последнего ошибка асимптотического приближения может оказаться сравнимой с НОВ.

В таблицах **1.1**, **1.2**, **1.3**, **1.4**, **1.5** приведены результаты сравнения точных значений НОВ и их асимптотических приближений для различных параметров моделей. Показывается, что новые асимптотические формулы существенно повышают точность аппроксимаций.

Раздел **1.3** посвящён последовательным d-гарантийным критериям различения гипотез $\mathrm{H}_0 : \theta \leqslant \theta_0$ и $\mathrm{H}_1 : \theta > \theta_0$.

В разделе **1.3.1** описывается универсальная последовательная d-гарантийная процедура (процедура первого перескока) и устанавливается её связь

[9] Володин И. Н., Новиков Ан. А. Асимптотика необходимого объёма выборки при d-гарантийном различении двух близких гипотез// *Известия ВУЗов. Математика* – 1983, № 11. – с. 59–66.

с последовательной процедурой вальдовского типа (**лемма 1.12**), у которой область продолжения наблюдений зависит от отношения условных правдоподобий при значениях параметра, принадлежащих различаемым гипотезам.

Для многих вероятностных моделей момент прекращения наблюдений универсальной процедуры может быть описан как момент первого выхода случайной суммы $S_n = \sum_1^n X_i,\ n = 1, 2, \ldots,$ за двусторонние границы параболического типа. В силу закона повторного логарифма, справедлива следующая

Лемма 1.11. *Пусть $X_1, X_2, \ldots$ — последовательность независимых одинаково распределенных случайных величин с конечным математическим ожиданием μ и конечной дисперсией σ^2. Тогда с вероятностью единица момент остановки*

$$\nu := \min\left\{n:\ S_n \leqslant a_{0n}\ \text{ или }\ a_{1n} \leqslant S_n\right\}\ <\ \infty,$$

если $a_{0n} < a_{1n},\ n \geqslant 1,$ и $a_{0n},\ a_{1n} = O(\sqrt{n}),\ n \to \infty$.

Для нормально-нормальной, бета-Бернулли и гамма-показательной моделей момент остановки универсальной d-гарантийной процедуры есть момент выхода за границы параболического типа — формула (**1.60**), **лемма 1.13** и **лемма 1.15**. Таким образом, в соответствии с леммой 1.11 для указанных вероятностных моделей момент остановки универсальной процедуры замкнут.

В нормально-нормальной модели момент остановки универсальной процедуры не только почти наверное конечен, но и имеет конечное математическое ожидание для значений параметра, отличных от границы между гипотезами.

Теорема 1.5. *Пусть $X_1, X_2, \ldots$ — независимые нормальные $\mathcal{N}(\theta, \sigma^2)$ случайные величины. Тогда для любого $\theta \neq \theta_0$ среднее значение момента остановки ν_{un} универсальной процедуры $\mathbf{E}_\theta\left[\nu_{un}\right] < \infty$.*

В граничной точке $\theta = \theta_0$ среднее значение ν_{un} бесконечно.

Теорема 1.6. *Пусть $X_1, X_2, \ldots$ — независимые нормальные $\mathcal{N}(\theta_0, \sigma^2)$ случайные величины. Если ограничения β_0, β_1 и среднее значение априорного распределения μ таковы, что $q = \max\{\Phi^{-1}(1-\beta_j),\ j = 0, 1\} > 1$ и $|\mu - \theta_0| \leqslant \sqrt{q^2 - 1}$, то среднее значение момента остановки универсальной d-гарантийной процедуры $\mathbf{E}_{\theta_0}\left[\nu_{un}\right] = \infty$.*

Высказывается предположение о бесконечности безусловного среднего ν_{un}, что подтверждается с помощью примеров, полученных методом стохастического моделирования (рис. **1.1**, таблицы **2.1**, **2.2**, **2.3**, **2.4**, **2.5**).

В разделе **1.3.2** обосновывается возможность применения усечённой универсальной процедуры, которая принудительно останавливается на каком-то фиксированном шаге.

В завершении раздела **1.3** для трёх рассмотренных вероятностных моделей описывается область продолжения наблюдений последовательной процеду-

ры на статистике вклада [10] $\sum_1^n \partial \ln f(x_i \,|\, \theta)/\partial\theta$. Устанавливается конечность безусловного среднего момента остановки ν_{sc} для этой процедуры.

Теорема 1.7. *Момент остановки ν_{sc} замкнут относительно безусловного распределения:* $\mathbf{P}\big(\nu_{sc} < \infty\big) = 1$, *и, кроме того, его математическое ожидание* $\mathbf{E}\nu_{sc} < \infty$.

Раздел **1.4** посвящён способам построения эмпирических аналогов d-гарантийных процедур. Пусть $\mathbf{X}^{(k)} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$ — результаты наблюдений в последовательности статистических экспериментов в рамках одной и той же вероятностной модели. Таким образом, $\mathbf{x}_j$ есть реализация случайного вектора с безусловным распределением (плотностью $f(x^{(n)}) = \int_\Theta \prod_1^n f(x^{(n)} \,|\, \theta)\, G(d\theta)$, где G — функция распределения ϑ).

Выписываются семейства достаточных статистик для вероятностной модели экспоненциального типа. В рамках модели N–N уточняется вид оценок максимального правдоподобия с учётом того, что решения уравнений правдоподобия не всегда попадают в область допустимых значений параметров.

Для моделей G–E и B–B устанавливается факт идентифицируемости параметров модели по безусловному распределению (**теорема 1.8**, **теорема 1.9**). Для описанных трёх моделей приведён вид информационных матриц Фишера, на основе которых в главе 2 строятся доверительные утверждения о параметрах этих моделей. Предлагаются способы построения оценок, подобных оценкам метода моментов (**лемма 1.16** для модели G–E, **лемма 1.17** для модели B–B).

В разделе **1.4.2** изучаются возможности применения непараметрических оценок плотности априорного распределения.

Теорема 1.10. *Предположим, что оценка $\widehat{g}_k(\theta\,;\, \mathbf{X}^{(k)})$ априорной плотности g такова, что при $k \to \infty$ для некоторой последовательности $v_k \to \infty$ расхождение в L_1 -метрике $\int_{\mathbb{R}^1} |\widehat{g}_k(\theta\,;\, \mathbf{X}^{(k)}) - g(\theta)|\, d\theta = O_{\mathbf{P}}(1/v_k)$ относительно маргинального (безусловного) распределения $\mathbf{X}^{(k)}$. Тогда при $k \to \infty$ относительно безусловного распределения $\mathbf{X}^{(k)}$*

$$\sup_{c \in Q_0} |\mathcal{R}_0(c\,;\, \widehat{g}_k) - \mathcal{R}_0(c\,;\, g)| \;=\; O_{\mathbf{P}}(1/v_k), \text{ где } Q_0 = \{c:\; F_\xi(c) > 0\}.$$

Обсуждаются способы выборы «окна» ядерной оценки. На конкретном примере с реальными данными демонстрируется преимущество выбора окна с учётом близости оценки функции распределения к эмпирической функции распределения статистики.

В разделе **2.1 главы 2** гарантийные статистические процедуры, разработанные в главе 1, применяются к задачам контроля качества и сравниваются по объёму выборок при различных параметрах вероятностной модели. Основная цель — изучить возможности применения последовательных схем контроля (в частности усечённой универсальной процедуры).

[10] Володин И. Н., Новиков Ан. А. Локальная асимптотическая эффективность последовательного критерия отношения вероятностей при гарантийном различении сложных гипотез// *ТВ и П* – 1998, № 2, с. 209–225.

В разделе **2.1.1** рассматривается нормально-нормальная модель при различных значениях входных параметров (таблицы **2.1**, **2.2**). Характеристики последовательных процедур (универсальной и на статистике вклада) находятся методом стохастического моделирования по большому числу репликаций. Делается вывод, что усечённая универсальная процедура может быть с успехом применена для задач контроля качества, т.к. она приводит к значительному сокращению среднего объёма выборки, сохраняя на приемлемом уровне надёжность статистических решений (предложение **2.1**). Этот же вывод подтверждают расчёты, основанные на данных реального производства (таблица **2.3**). Кроме того, из этой таблицы видно, что наибольшее сокращение объёма испытаний происходит во время инспекции кондиционной продукции, что весьма полезно для схем контроля с разрушением. В то же время, процедура на статистике вклада не позволяет контролировать на заданном уровне надёжность статистических решений.

К аналогичным результатам приводит рассмотрение моделей гамма-показательная (раздел **2.1.2**, таблица **2.4**, предложение **2.2**) и бета-Бернулли (раздел **2.1.3**, таблица **2.5**, предложение **2.3**)

В разделе **2.2** методика построения d-гарантийных процедур применяется к проблеме множественного тестирования, в частности к проблеме выделения из большого числа генов тех из них, которые имеют повышенную и пониженную экспрессию среди заболевших пациентов.

Вначале даётся обзор существующих подходов к определению характеристик надёжности статистического вывода при множественном тестировании и их связь с функцией d-риска.

В разделе **2.2.1** рассматривается задача выявления генов с изменённой экспрессией у пациентов с онкологическим заболеванием по значениям двухвыборочной статистики Стьюдента T (всего имелось $M = 6033$ гена; данные взяты из монографии [8]). Строятся две модели (согласующиеся с данными), в которых наблюдение T трактуется как реализация нормальной случайной величины с единичной дисперсией и случайным средним значением ϑ, характеризующим разность экспрессий в двух экспериментальных группах пациентов. В обеих моделях априорное распределение ϑ есть смесь распределения, сосредоточенного с вероятностью $1-\pi$ в точке $\theta = 0$, с нормальным распределением или со смесью двух нормальных распределений:

$$\mathbf{P}\big(\vartheta < \theta\big) = (1-\pi)\mathbb{I}_0(\theta) + \pi G(\theta).$$

где $\mathbb{I}_0(\theta)$ — индикаторная функция множества $\theta \in (0,\infty)$, π — доля генов с изменённым уровнем экспрессии; функция распределения $G(\theta) = \Phi((\theta-\mu)/\tau)$ или $\pi G(\theta) = \pi_1\Phi((\theta+\mu_1)/\tau_1) + \pi_2\Phi((\theta-\mu_2)/\tau_2)$ с параметрами $\tau, \tau_1, \tau_2 > 0, \mu, \mu_1, \mu_2 \geqslant 0$. Решается задача выделения генов с изменённой экспрессией (нулевая гипотеза $\mathrm{H}_0 : \theta = 0$ при двусторонней альтернативе $\mathrm{H}_1 : \theta \neq 0$) и

задача выделения генов с повышенной экспрессией (нулевая гипотеза $\mathrm{H}_0 : \theta \leqslant 0$ при альтернативе $\mathrm{H}_1 : \theta > 0$).

Для построения d-гарантийного критерия в задаче выделения генов с изменённой экспрессией сначала доказывается

Теорема 2.1. *Пусть справедлива модель с нормальным распределением T и априорным распределением, представимым в виде смеси с нормальным распределением G; $\Pi_0(t)$ — апостериорная вероятность события $\vartheta = 0$. Тогда*

(i) функция $\Pi_0(t)$, $t \in \mathbb{R}^1$, имеет единственный локальный максимум в точке $t = t^* = -\mu/(\gamma_n \tau^2)$;

(ii) функция $\Pi_0(t)$, $t \in \mathbb{R}^1$, симметрична около точки t^*, т.е. $\Pi_0(t+t^*) = \Pi_0(-t+t^*)$;

(iii) неравенство $\Pi_0(t) < C$ (для $C \leqslant \Pi_0(t^*)$) выполняется тогда и только тогда, когда $t^* - c \leqslant t \leqslant t^* + c$, где $c \geqslant 0$ и $\Pi_0(t^* - c) = \Pi_0(t^* + c) = C$.

Проведённые численные эксперименты показывают (таблица **2.1**, предложение **2.4**), что процедура Бенжамини–Хочберга более консервативна, чем оптимальный d-гарантийный тест — она почти вдвое реже отвергает нулевую гипотезу. Кроме того, здесь возможно построение процедуры, гарантирующей величину средних потерь при принятии нулевой гипотезы (аналог характеристики pFNR), а также процедуры, гарантирующей обе величины средних потерь (при соответствующем увеличении числа обследуемых пациентов).

Обсуждается проблема различения трёх гипотез (выделение генов с пониженной, повышенной или с «нормальной» экспрессией у пациентов экспериментальной группы). В этой ситуации вместо функции d-риска рассматривается функция надёжности, т.е. условная вероятность справедливости той или иной гипотезы, если принято решение в её пользу. Численно устанавливается вид минимаксной процедуры, позволяющий предположить, что для этой процедуры надёжность всех трёх решений совпадает (замечание 15).

В разделе **2.2** даётся общая схема построения оптимального d-гарантийного критерия в задаче сравнения двух групп. Далее, в разделе **2.2.1** эта схема применяется к ситуации, когда распределения в обеих группах нормальны со случайными средними значениями и фиксированными дисперсиями; распределения средних значений в группах также предполагаются нормальными. Эмпирическим путём было замечено, что

при различении гипотез $H_0 : \theta \leqslant 0$ и $H_1 : \theta > 0$ (а также гипотез $H_0 : \theta = 0$ и $H_1 : \theta \neq 0$) тестовая статистики оптимального критерия зависит от линейной комбинации выборочных средних $\overline{X}, \overline{Y}$ вида

$$S = \overline{X} - \left(1 + \frac{\sigma_x^2}{n\gamma_0^2}\right) \overline{Y},$$

где n и σ_x^2 — объём выборки и дисперсия наблюдений в контрольной группе, τ_0^2 — дисперсия распределения среднего значения в контрольной группе. Опти-

мальный критерий принимает гипотезу H_0, *если* $S > C$ *(или* $C_1 < S < C_2$ *) с соответствующим образом подобранной константой* C *(константами* C_1, C_2 *).*

Заключение

В диссертационной работе исследованы свойства универсальной последовательной процедуры различения двух односторонних гипотез в рамках d-апостериорного подхода. В частности, показано, что для вероятностных моделей, для которых область продолжения наблюдений имеет границы параболического типа, момент остановки этой процедуры замкнут. К таким моделям относятся нормально-нормальная, бета-Бернулли и гамма-показательная модели. Доказано, что математическое ожидание момента остановки универсальной последовательной процедуры для нормально-нормальной модели при истинном значении параметра, отличном от границы между гипотезами, конечно; в противном случае среднее значение момента остановки этой процедуры равно бесконечности.

Изучены возможности применения, а также свойства усечённой универсальной последовательной процедуры. Численными методами показана её гарантийность и высокая эффективность по среднему объёму наблюдений.

Для двух частных вероятностных моделей, найдены новые асимптотические формулы для необходимого объёма выборки, значительно повышающие точность аппроксимации.

Методика d-апостериорного подхода применена к задачам множественного тестирования, что позволило контролировать на заданном уровне не только долю ошибок среди отвергнутых нулевых гипотез, но и долю ошибок среди решений в пользу нулевой гипотезы. Кроме того, эта методика даёт возможность управления наблюдениями. Разработана общая схема построения процедур сравнения при множественном тестировании.

Работы автора по теме диссертации

Научные статьи, опубликованные в журналах ВАК, WoS, SCOPUS

[1] Симушкин Д. С. Сравнительный анализ по объёму выборки двух последовательных d-гарантийных процедур// *Обозрение Прикладной и Промышленной Математики* – 2011, т. **18**, № 1. – с. 91—93.

[2] Simushkin D. S. Empirical estimation of d-risks at distinguishing one-sided hypotheses// *Lobachevskii Journal of Mathematics* – 2016, Vol. **37**, № 4. – p. 509—514.

[3] Simushkin D. S., Simushkin S. V., Volodin I. N. D-guaranteed discrimination of statistical hypotheses: review of results and unsolved probems// *Journal of Mathematical Science* – 2018, Vol. **228**, № 5, February. – pp. 543 –565

[4] Simushkin D. S., Simushkin S. V., Volodin I. N. On the d-posterior approach to the multiple testing problem// *Journal of Statistical Computation and Simulation* – 2019

[5] Simushkin D. S. Asymptotic of the necessary sample size in the two hypotheses discrimination problem// *Lobachevskii Journal of Mathematics* – 2019, Vol. **40**, № 2. – p. (принята к печати)

Тезисы докладов на научных конференциях

[6] Simushkin D. S., Volodin I. N. FDR is the d-risk// Abstracts Communications " SysPatho Workshop «Systems Biology and Medicine»". St.Petersburg. – 2012, p. 88–89

[7] Симушкин Д. С. О точности эмпирических оценок d-апостериорного риска// *Обозрение Прикладной и Промышленной Математики* – 2013, т. **20**, № 2. – с. 153.

[8] Simushkin D. S., Simushkin S. V., Volodin I. N. pFDR & d-risk: Large-Scale Inference for Genes Expression Data// Abstracts Comm. 11th Internat. Vilnius Conf. on Probab. Theory and Mathem. Statist. – 2014. – p. 227

[9] Симушкин Д. С. Процедуры различения многих гипотез при множественном тестировании// *Материалы межд.конф. по алгебре, анализу и геометрии* – Казань: Изд-во Академии наук РТ, 2016. – с. 313–314.