

КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

На правах рукописи

СИМУШКИН ДМИТРИЙ СЕРГЕЕВИЧ

**СТАТИСТИЧЕСКИЕ КРИТЕРИИ С ОГРАНИЧЕНИЯМИ
НА d-РИСКИ**

Специальность: 01.01.05 – теория вероятностей и математическая
статистика

Диссертация
на соискание учёной степени
кандидата физико-математических наук

Научный руководитель –
доктор физико-математических наук,
профессор
Володин Игорь Николаевич

Казань–2020

Оглавление

Список обозначений и сокращений	4
Введение	5
Глава 1. Статистические критерии проверки параметрических гипотез с ограничениями на d-риски	18
1.1. Вероятностная модель статистического эксперимента	19
1.2. D-гарантийные критерии на фиксированном числе наблюдений .	26
1.2.1. Асимптотика необходимого объёма выборки в схеме с жёсткими ограничениями на d-риски	39
1.2.2. Асимптотика необходимого объёма выборки в схеме стягивающегося априори	52
1.3. Последовательные d-гарантийные критерии	60
1.3.1. Процедура первого перескока	60
1.3.2. Усечённая процедура первого перескока	63
1.3.3. Последовательный критерий на статистике вклада	64
1.4. Эмпирические аналоги d-гарантийных процедур	74
1.4.1. Эмпирические оценки параметров вероятностной модели . .	75
1.4.2. Непараметрические оценки плотности априорного распределения	86
Глава 2. D-апостериорный подход к проблемам контроля каче- ства и множественного тестирования генов	91
2.1. Статистический контроль качества	92
2.1.1. Выбор процедуры контроля качества в рамках модели N–N	94
2.1.2. Приёмочный контроль в рамках модели G–E	100
2.1.3. Контроль по альтернативному признаку (модель B–B) . . .	102
2.2. Введение в проблему множественного тестирования	104
2.2.1. Тестирование экспрессии генов	108
2.2.2. Вероятностная модель с тремя параметрами	110
2.2.3. Модель с шестью параметрами	117

2.3. Общая модель сравнения двух групп	120
2.3.1. Модель N–N	122
Заключение	126
Список иллюстраций	127
Список таблиц	128
Список литературы	129

Список обозначений и сокращений

\mathbf{P}_θ	— распределение наблюдений при истинном значении параметра θ
\mathbf{P}	— совместное распределение наблюдений и случайного параметра
\mathbf{E}_θ	— математического ожидание относительно распределения при истинном значении параметра θ
\mathbf{E}	— математического ожидание относительно совместного распределения наблюдений и случайного параметра
\mathbb{I}_A	— индикаторная функция события A
\mathbf{I}_0	— информация по Фишеру в граничной точке θ_0
$\mathcal{N}(\mu, \sigma^2)$	— нормальный закон со средним μ и дисперсией σ^2
Φ	— функция распределения нормального $\mathcal{N}(0, 1)$ закона
$\mathbf{B}(p, q)$	— бета-функция Эйлера
$\mathcal{B}(p, q)$	— бета-закон с параметрами p, q
$\mathbb{B}(\cdot; p, q)$	— функция распределения бета-закона $\mathcal{B}(p, q)$
$\Gamma(p)$	— гамма-функция
$\mathcal{G}(\lambda, a)$	— гамма-закон с параметром формы λ и интенсивностью a
$\mathbb{G}(\cdot; \lambda, a)$	— функция распределения гамма-закона $\mathcal{G}(\lambda, a)$
$\Pi_0 = \mathbf{P}(\vartheta \in \Theta_0)$	— априорная вероятность нулевой гипотезы

Введение

Актуальность темы исследования. Начиная с 80-х годов XX века, с подачи Л.Н.Большева, усилиями И.Н.Володина и его учеников (см. [6], [21], [12]) получил развитие так называемый d -апостериорный подход к проблеме гарантийности статистического вывода. В этом подходе риск любого статистического правила 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. Качество любого статистического правила должно учитывать априорную информацию о распределении этой характеристики.

Идее d -апостериорного подхода к проблеме гарантийного различения гипотез предшествовал прорыв в байесовской теории статистического вывода, наметившийся в середине XX века. В своей статье [44] Дж.Нейман обратил внимание на широкий класс статистических задач, в которых выводной параметр является реализацией случайной величины, и проблема состоит только в спецификации его априорного распределения. Заслуга Г.Роббинса, стоявшего у истоков этого прорыва, состоит в том, что такие задачи всегда сопровождаются наличием большого архива данных предыдущих исследований аналогичных объектов, которые поставляют информацию об априорном распределении параметра. Г.Роббинс [46] привёл несколько примеров построения эмпирической байесовской оценки на основе архивных данных, не требующих описания вероятностной модели для априорного распределения.

Отметим, что ещё до работ Г.Роббинса, С.Н.Бернштейн [2] критикует применение стандартного (не байесовского) подхода к проблеме гарантийности статистического контроля качества. Ограничения на риски потребителя и

производителя не отвечают существу задачи — в контроле качества надо гарантировать не долю ошибочно принятой некондиционной продукции, а долю плохой продукции, полученной потребителем. Стандартный подход игнорирует реально существующее априорное распределение контролируемого параметра. Даже в том случае, когда априорное распределение параметра неизвестно, в задачах контроля качества всегда существует большой массив данных предыдущих инспекций аналогичных объектов, на основе которого можно извлечь информацию об априорном распределении, следуя идеям Г. Роббинса.

Такого рода подход к понятию риска статистического решения специфичен для любого статистического исследования, в котором проводится анализ последовательно поступающих массивов данных по наблюдениям однотипных объектов: данных медицинских обследований, социологических обследований по малым областям, данных биочипов при генетических исследованиях. Для большинства таких задач байесовский риск является слишком грубой характеристикой величины средних потерь. Усреднять потери следует среди тех статистических исследований, которые закончились принятием одного и того же конкретного решения. О целесообразности рассмотрения такого рода потерь говорил также С.Р. Рао (1952 г.) в своей монографии [45], рассматривая байесовские правила классификации.

Некоторые общие вопросы оптимального статистического вывода в рамках d -апостериорного подхода были решены И.Н. Володиным и его учениками (см. обзоры [12], [68]). В частности, была предложена универсальная последовательная гарантийная процедура, которая останавливается, когда апостериорный риск байесовского решения удовлетворяет заданным ограничениям. К сожалению, вопросы замкнутости и конечности среднего значения её момента остановки подробно не рассматривались. Кроме того, были найдены асимптотические формулы для необходимого объёма выборки при различении двух односторонних параметрических гипотез.

В статье [53, В. Sorić] рассматривались некоторые способы анализа близкой к понятию d -риска характеристики, названной FDP — от английского false discovery proposition (доля ложных обнаружений), т.е. относительная доля ошибочно отклонённых нулевых гипотез в большой группе однотипных ста-

статистических задач. Предложенный в [28, J. Benjamini, J. Hochberg] алгоритм позволяет контролировать среднюю долю ложных обнаружений — так называемый показатель FDR (false discovery rate). Несомненным преимуществом этого алгоритма является то, что контроль FDR возможен вне байесовской теории без каких-либо предположений относительно распределения количества ложных и справедливых нулевых гипотез. К сожалению, показатель FDR не вполне отражает ожидания исследователя о величине средней ошибки. При вычислении FDR показатель FDP полагается равным нулю, если отсутствуют эксперименты с отклонёнными нулевыми гипотезами, что приводит к значительному уменьшению средней доли ошибок в ситуациях, когда количество верных нулевых гипотез близко к 100%. Вместо FDR был предложен показатель pFDR, равный условному среднему значению FDP при условии, что среди всей совокупности экспериментов обязательно найдётся хоть один эксперимент с отклонённой нулевой гипотезой.

В байесовской постановке показатель pFDR (см. [56]) полностью совпадает с понятием d-риска от решения в пользу альтернативы. В работе J.D. Storey [57] для показателя pFDR доказано утверждение, подобное основной лемме работы [21] (аналог леммы Неймана-Пирсона). Подробное обсуждение этой концепции можно найти в монографии В. Efron [34], где разрабатываются гарантийные процедуры составного статистического решения на примере данных биочипов (microarray data). Таким образом, отрывается новая перспективная область применения d-апостериорного подхода к анализу обширных генетических экспериментов.

Цель и задачи исследования. Целью диссертационной работы является сравнительный анализ ранее предложенных статистических процедур различения двух односторонних гипотез $H_0 : \theta \in \Theta_0 = (-\infty, \theta_0]$ и $H_1 : \theta \in \Theta_1 = (\theta_0, \infty)$ о действительном параметре θ в рамках d-апостериорного подхода. А именно: исследование свойств момента останова универсальной процедуры (замкнутость момента останова и конечность его математического ожидания) и сравнение этой процедуры с последовательной процедурой на статистике вклада и d-гарантийной процедурой с фиксированным числом на-

блюдений; уточнение асимптотических формул необходимого объёма выборки; разработка методики применения d-апостериорного подхода к задачам множественного тестирования.

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.

Краткое содержание диссертации. В **Главе 1** изучаются свойства различных методов построения d-гарантийных критериев, основанных как на фиксированном объёме выборки, так и в рамках последовательных схем получения выборочных данных.

В разделе **1.1** описываются основные положения теории принятия решений с ограничениями на d-апостериорные вероятности ошибок. Пусть требуется проверить гипотезу $H_0 : \theta \in \Theta_0$ о параметре θ , индексирующем распределение \mathbf{P}_θ наблюдаемой случайной величины X , которое можно описать с помощью плотности $f(\cdot | \theta)$ относительно некоторой сигма-конечной меры μ . Предполагается, что значение θ есть реализация случайной величины ϑ с некоторой функцией распределения G (плотностью g относительно меры γ). Если решение d_0 в пользу H_0 или решение d_1 в пользу альтернативы $H_1 : \theta \notin \Theta_0$ принимается посредством решающей функции δ на основе последовательности наблюдений $X^{(\nu)} = (X_1, \dots, X_\nu)$ с моментом остановки ν , то d-риск 1-го рода δ определяется как условная вероятность

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

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

Лемма 1.1. Пусть δ — некоторая решающая функция в задаче различения двух гипотез $H_0 : \theta \in \Theta_0$, $H_1 : \theta \notin \Theta_0$, $\Pi_0 = \mathbf{P}(\vartheta \in \Theta_0)$ — априорная

вероятность Θ_0 , $\Psi(d_0) = \mathbf{P}(\delta = d_0)$ — 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) $n = n^*$, гарантирующего заданные ограничения на обе d -апостериорные вероятности ошибок:

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

Рассматривается задача различения гипотез $H_0 : \theta \leq \theta_0$ и $H_1 : \theta > \theta_0$ 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 В–В с бернуллиевским распределением наблюдений (θ — вероятность «успеха») и бета-распределением θ .

Если в модели N–N граница между гипотезами совпадает с априорным средним, то величину НОВ можно найти точно.

Теорема 1.1. Для задачи различения гипотез $H_0 : \theta \leq \mu$ и $H_1 : \theta > \mu$ в рамках модели N–N, где μ — априорное среднее неизвестного параметра, 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 $\tau^2 \rightarrow \infty$ при различении гипотез $H_0 : \theta \leq \theta_0$ и $H_1 : \theta > \theta_0$ необходимый объём выборки $n^* \rightarrow 1$.

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 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 $\mathcal{G}(n, 1)$ имеет асимптотическое (при $n \rightarrow \infty$) представление

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

В разделе **1.2.1** изучается асимптотика НОВ n^* при различении гипотез $H_0 : \theta \leq \theta_0$, $H_1 : \theta > \theta_0$ в ситуации, когда ограничения $\beta_0, \beta_1 \rightarrow 0$. В **леммах 1.6** и **1.7** даются упрощённые представления для асимптотических формул из [58]. 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, $\rho = \beta_1/\beta_0$, $W(c) = \phi(c) + c\Phi(c)$, $c \in (-\infty, \infty)$, $\Pi_0 = \Phi((\theta_0 - \mu)/\tau)$ — априорная вероятность справедливости нулевой гипотезы, $g_0 = \tau^{-1}\phi((\theta_0 - \mu)/\tau)$ — значение априорной плотности в граничной точке. Выберем c_0 как решение уравнения $W(c_0)(\Pi_0(1 + \rho) - \rho) = c_0\Pi_0$ и положим $\Delta_0 = \rho - (1 + \rho)\Pi_0$, $Z = \Pi_0 + \Delta_0\Phi(c_0)$.

Теорема 1.3. Пусть $\beta_0, \beta_1 \rightarrow 0$ 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 $[a]$ — наименьшее целое число, не меньше a , и

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

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

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 с априорной гамма-плотностью при различении гипотез $H_0: \theta \leq \theta_0$, $H_1: \theta > \theta_0$ и ограничениях $\beta_1 = \rho\beta_0 \rightarrow 0$ необходимый объём выборки

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

где

$$\lim_{\beta_0 \rightarrow 0} v_1(\beta_0) = \frac{V_1}{Z},$$

$$V_1 = -\frac{\Delta_0}{6} (3a\theta_0 + 1 - 3\lambda)\phi(c_0) + \rho(\theta_0 g_0 - (a\theta_0 - \lambda)(1 - \Pi_0))c_0 - \\ - \theta_0 g_0 (1 + \rho)\Phi(-c_0)c_0 + \Delta_0(a\theta_0 - \lambda)\Phi(-c_0)c_0.$$

В разделе 1.2.2 изучаются возможности применения к описанным трём вероятностным моделям аппроксимации НОВ n^* в схеме стягивающегося априори, полученной в [8]. В этой схеме предполагается, что априорная плотность может быть представлена в виде $g(\theta) = \frac{1}{\tau} \tilde{g}((\theta - \theta_0)/\tau; \tau)$, где функция $\tilde{g}(\cdot; \tau) \rightarrow \tilde{g}(\cdot; 0)$ при $\tau \rightarrow 0$.

В лемме 1.8 устанавливается, что для нормально-нормальной модели асимптотические формулы [8] дают способ вычисления точного значения НОВ n^* . Для вероятностных моделей G–E и В–В функция \tilde{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 $H_0 : \theta \leq \theta_0$ и $H_1 : \theta > \theta_0$. В работе [7] была предложена универсальная (пригодная для любой статистической проблемы) последовательная d-гарантийная процедура. В проблеме различения двух указанных гипотез продолжение наблюдений этой процедуры на шаге n после получения выборки $x^{(n)}$ определяется через апостериорную вероятность справедливости нулевой гипотезы посредством неравенств

$$\beta_1 < \mathbf{P}\{\vartheta \leq \theta_0 \mid x^{(n)}\} < 1 - \beta_0.$$

Для многих вероятностных моделей момент прекращения наблюдений универсальной процедуры может быть описан как момент первого выхода случайной суммы $S_n = \sum_1^n X_i$, $n = 1, 2, \dots$, за двусторонние границы параболического типа. Замкнутость этого момента можно проверить с помощью закона повторного логарифма. Справедлива следующая

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

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

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

В разделе **1.3.1** описывается универсальная последовательная d-гарантийная процедура (процедура первого перескока) и устанавливается её связь с последовательно процедурой вальдовского типа (**лемма 1.12**), у которой область продолжения наблюдений зависит от отношения условных правдоподобий при значениях параметра, принадлежащих различаемым гипотезам.

Для нормально-нормальной, гамма-показательной и бета-Бернулли моделей момент остановки универсальной d-гарантийной процедуры есть момент

выхода за границы параболического типа — формула (1.60), лемма 1.13 и лемма 1.14. 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, \dots — независимые нормальные $\mathcal{N}(\theta, \sigma^2)$ случайные величины. Тогда для любого $\theta \neq \theta_0$ среднее значение момента остановки ν_{un} универсальной процедуры $\mathbf{E}_\theta[\nu_{un}] < \infty$.

В граничной точке $\theta = \theta_0$ математическое ожидание момента остановки ν_{un} бесконечно.

Теорема 1.6. Пусть X_1, X_2, \dots — независимые нормальные $\mathcal{N}(\theta_0, \sigma^2)$ случайные величины. Если априорное среднее $|\mu - \theta_0| \leq \sqrt{q^2 - 1}$, где $q = \min\{\Phi^{-1}(1 - \beta_j), j = 0, 1\} > 1$, то среднее значение момента остановки универсальной d -гарантийной процедуры $\mathbf{E}_{\theta_0}[\nu_{un}] = \infty$.

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

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

Раздел 1.4 посвящён способам построения эмпирических аналогов d -гарантийных процедур. Пусть $\mathbf{X}^{(k)} = (\mathbf{x}_1, \dots, \mathbf{x}_k)$ — результаты наблюдений в последовательности статистических экспериментов в рамках одной и той же

вероятностной модели. Таким образом, \mathbf{x}_j есть реализация случайного вектора с плотностью $f(x^{(n)}) = \int_{\Theta} \prod_1^n f(x^{(n)} | \theta) G(d\theta)$, где G — функция распределения ϑ (относительно меры μ).

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

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

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 **Главе 2** d-гарантийные статистические процедуры, разработанные в первой главе, применяются к задачам контроля качества и сравниваются по объёму выборок при различных параметрах вероятностной модели. Основная цель — показать преимущества последовательных схем и выяснить границы их применимости. Кроме того, методика построения d-гарантийных процедур

применяется к проблеме множественного тестирования, в частности к проблеме выделения из большого числа генов тех из них, для которых характерно изменение экспрессии в группе заболевших пациентов. Показывается, что в рамках d -апостериорного подхода можно контролировать оптимальным образом не только d -риск первого рода (байесовский аналог $pFDR$), но и d -риск второго рода (аналог $pFNR$), а также строить процедуры различения более двух гипотез.

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 статистического вывода (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 «плохих» генов.

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

В разделе **2.2.1** рассматривается задача выявления генов с изменённой экспрессией (из общей совокупности $M = 6033$ гена) у пациентов с онкологическим заболеванием по значениям двухвыборочной статистики Стьюдента

T (данные взяты из монографии [34]). 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 Π_0 в точке $\theta = 0$, и нормального распределения или смеси двух нормальных распределений со средними разного знака:

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

где $\mathbb{I}_0(\theta)$ — индикаторная функция множества $\theta \in (0, \infty)$, π ($= \pi_1$) — 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 $\tau, \tau_1, \tau_2 > 0, \mu, \mu_1, \mu_2 \geq 0$. Решается задача выделения генов с изменённой экспрессией (нулевая гипотеза $H_0 : \theta = 0$ при двусторонней альтернативе $H_1 : \theta \neq 0$) и задача выделения генов повышенной экспрессией (нулевая гипотеза $H_0 : \theta \leq 0$ при альтернативе $H_1 : \theta > 0$).

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 $C \leq \Pi_0(t^*)$) выполняется тогда и только тогда, когда $t^* - c \leq t \leq t^* + c$, где $c \geq 0$ 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). 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 $H_0 : \theta \leq 0$ и $H_1 : \theta > 0$ (а также гипотез $H_0 : \theta = 0$ и $H_1 : \theta \neq 0$) тестовая статистика оптимального критерия зависит от линейной комбинации выборочных средних \bar{X}, \bar{Y} вида

$$S = \bar{X} - \left(1 + \frac{\sigma_x^2}{n\tau_0^2}\right) \bar{Y},$$

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

Глава 1

Статистические критерии проверки параметрических гипотез с ограничениями на d-риски

В этой главе изучаются различные методы построения d-гарантийных критериев, основанных как на фиксированном объёме выборки, так и в рамках последовательных схем получения выборочных данных.

Для фиксированного числа наблюдений в теории d-гарантийного различения гипотез существует аналог классического критерия Неймана–Пирсона, который, гарантируя заданное ограничение на d-риск первого рода, минимизирует значение d-риска второго рода (см. [21]). Естественно, такой критерий минимизирует и объём выборки, необходимый для различения двух гипотез с заданными ограничениями β_0 и β_1 на соответствующие d-риски первого и второго рода. Это так называемый «необходимый объём выборки» (в дальнейшем будет использоваться аббревиатура НОВ).

В диссертации анализируются два известных ([58, теорема 2.2] и [8]) метода получения асимптотических формул для НОВ в задаче d-гарантийного различения односторонних параметрических гипотез. Полученные ранее асимптотические формулы уточняются в рамках конкретных, наиболее распространенных в байесовской теории вероятностных моделей выборочных данных и априорных сведений.

Другое, не менее важное для теории и практики статистического вывода, достоинство конкретных формул НОВ состоит в использовании их как эталона для определения эффективности последовательных гарантийных критериев по среднему объёму наблюдений. В этой главе с этой точки зрения анализируются два последовательных d-гарантийных критерия: последовательный критерий «первого перескока» [7, 10] (так называемый универсальный последовательный критерий) и последовательный локально эффективный критерий, основанный на статистике вклада [10]. Различные характеристики этих критериев, связанные с объёмом испытаний, исследуются в рамках трёх вероятностных моделей: нормальное (ϑ, σ^2) распределение наблюдаемой случайной величини-

ны и нормальное априорное распределение ϑ при фиксированном значении σ^2 ; показательное распределение наблюдений с параметром интенсивности ϑ и априорное гамма-распределение ϑ ; выбор в схеме Бернулли с вероятностью успешного испытания ϑ и априорное бета-распределение ϑ . Показывается, что для этих моделей последовательные схемы замкнуты, т.е. останавливаются с вероятностью единица на конечном числе испытаний, и имеют для почти всех значений параметров θ конечное математическое ожидание. Приводятся доводы в пользу того, что момент остановки универсальной последовательной процедуры имеет бесконечное безусловное среднее значение. Эти доводы, кроме теоретических соображений, подкрепляются данными характеристик критериев, полученных методом статистического моделирования.

1.1 Вероятностная модель статистического эксперимента

Прежде чем перейти к рассмотрению конкретных задач по d-гарантийному различению статистических гипотез, опишем основные положения общей теории d-апостериорного подхода к принятию решений в духе [7].

Рассмотрим ряд статистических экспериментов $\mathcal{E}_1, \mathcal{E}_2, \dots$. Предположим, что в эксперименте с номером k ($= 1, 2, \dots$) наблюдается последовательность X_1, X_2, \dots независимых (в этом эксперименте) копий случайной величины с распределением \mathbb{P}_k на измеримом пространстве $(\mathbb{X}, \mathcal{F})$. Распределение \mathbb{P}_k полностью определяется значением некоторого неизвестного параметра $\theta \in \Theta$. Другими словами, имеется семейство вероятностных распределений $\mathcal{P} = \{\mathbf{P}_\theta, \theta \in \Theta\}$ такое, что $\mathbb{P}_k = \mathbf{P}_{\theta_k}$ для некоторого $\theta_k \in \Theta$ и независимость X_1, X_2, \dots подразумевается относительно распределения \mathbf{P}_{θ_k} , т.е.

$$\mathbf{P}_{\theta_k}(X_1 \in A_1, \dots, X_n \in A_n) = \prod_{i=1}^n \mathbf{P}_{\theta_k}(X_i \in A_i)$$

для любых событий $A_1, \dots, A_n \in \mathcal{F}$ и любого $n \geq 1$.

Операцию вычисления математического ожидания относительно распределения \mathbf{P}_θ будем обозначать через \mathbf{E}_θ .

На практике распределение \mathbf{P}_θ описывается с помощью функции плотности $f(x | \theta), x \in \mathbb{X}$, относительно некоторой сигма-конечной меры μ на

$(\mathbb{X}, \mathcal{F})$. В дальнейшем будем предполагать, что функция $f(x | \theta)$, $x \in \mathbb{X}$, $\theta \in \Theta$, измерима относительно борелевской сигма-алгебры на \mathbb{R}_+^1 и сигма-алгебры $\mathcal{T} \otimes \mathcal{F}$ на $\Theta \times \mathbb{X}$, где \mathcal{T} — сигма-алгебра измеримых подмножеств Θ .

В рамках байесовской парадигмы предполагается, что значения $\theta_1, \theta_2, \dots$ параметра θ (определяющие распределения наблюдений в соответствующих экспериментах $\mathcal{E}_1, \mathcal{E}_2, \dots$) представляют собой независимые реализации случайной величины ϑ с априорным распределением Π на измеримом пространстве (Θ, \mathcal{T}) . Пусть $g(\theta)$, $\theta \in \Theta$, — плотность Π (априорная плотность) по некоторой мере γ на \mathcal{T} .

Априорное распределение считается полностью известным или же его можно оценить по архивным данным (см. раздел 1.4).

Распределение выборочного вектора $X^{(n)} = (X_1, \dots, X_n)$ с фиксированным числом наблюдений $n \geq 1$ при заданном параметре $\theta \in \Theta$ описывается плотностью

$$f_n(x^{(n)} | \theta) = \prod_1^n f(x_i | \theta), \quad \mathbf{x}^{(n)} \in \mathbb{X}^n,$$

относительно произведения мер $\mu_n = \otimes_1^n \mu$ на прямом произведении измеримых пространств $(\mathbb{X}^n, \mathcal{F}^n)$. Здесь, как и всюду в дальнейшем, символ \mathcal{F}^n используется для обозначения тензорного произведения сигма-алгебр (вместо более громоздких символов типа $\mathcal{F}^{\otimes n}$ или $\otimes_1^n \mathcal{F}$).

Совместное распределение вектора $(\vartheta, X^{(n)})$ можно задать плотностью (относительно меры $\gamma \otimes \mu_n$ на $(\Theta \times \mathbb{X}^n, \mathcal{T} \otimes \mathcal{F}^n)$)

$$f_n(x^{(n)} | \theta) g(\theta), \quad \theta \in \Theta, \quad x^{(n)} \in \mathbb{X}^n.$$

Набор (при $n = 1, 2, \dots$) этих конечномерных распределений задаёт совместное распределение ϑ и бесконечной последовательности наблюдений X_1, X_2, \dots . Операции вычисления математического ожидания и вероятности относительно совместного распределения $(\vartheta, X_1, X_2, \dots)$ будем обозначать символами **E** и **P** соответственно.

Очевидно, маргинальное распределение ϑ совпадает с априорным распределением Π . В свою очередь, плотность маргинального распределения $X^{(n)}$ (в байесовской статистике его ещё называют безусловным распределе-

нием) при каждом фиксированном $x^{(n)} \in \mathbb{X}^n$ вычисляется по формуле

$$f_n(x^{(n)}) = \int_{\Theta} f_n(x^{(n)} | \theta) g(\theta) \gamma(d\theta).$$

Договоримся, что прописные символы ϑ, X внутри соответствующих операций будут сигнализировать о тех случайных величинах, относительно которых вычисляются среднее значение или вероятность события. Строчные символы θ, x зарезервированы для конкретных значений этих случайных величин. В соответствии с этой договорённостью, безусловную плотность можно записать в виде

$$f_n(x^{(n)}) = \mathbf{E}[f_n(x^{(n)} | \vartheta)], \quad x^{(n)} \in \mathbb{X}^n.$$

Заметим, что плотность $f_n(x^{(n)} | \theta)$ есть теперь вариант плотности условного распределения $X^{(n)}$ при фиксированном значении $\vartheta = \theta$ (условного распределения на $\mathcal{T} \otimes \mathcal{F}^n$ относительно сигма-алгебры $\mathcal{T} \times \mathbb{X}^n$):

$$\mathbf{P}(X^{(n)} \in A | \vartheta) \Big|_{\vartheta=\theta} = \int_{x^{(n)} \in A} f_n(x^{(n)} | \theta) d\mu_n = \mathbf{P}_{\theta}(X^{(n)} \in A).$$

Таким образом, в соответствии с формулой полного математического ожидания

$$\mathbf{P}\{X^{(n)} \in A\} = \mathbf{E}[\mathbf{P}_{\vartheta}\{X^{(n)} \in A\}] \quad \text{и} \quad \mathbf{E}[h(X^{(n)})] = \mathbf{E}[\mathbf{E}_{\vartheta}(h(X^{(n)}))]$$

для любого события $A \in \mathcal{F}^n$ и любой неотрицательной измеримой функции h на выборочном пространстве \mathbb{X}^n .

Апостериорная вероятность любого измеримого события $B \subset \Theta$ после полученного результата наблюдений $X^{(n)} = x^{(n)}$ равна

$$\mathbf{P}(\vartheta \in B | x^{(n)}) = \frac{1}{f_n(x^{(n)})} \int_B f_n(x^{(n)} | \theta) \Pi(d\theta) = \int_B g(\theta | x^{(n)}) \gamma(d\theta), \quad (1.1)$$

где $g(\theta | x^{(n)}) = f_n(x^{(n)} | \theta)g(\theta)/f_n(x^{(n)})$, $\theta \in \Theta$, есть апостериорная плотность ϑ , т.е. плотность условного распределения на $(\Theta \times \mathbb{X}^n, \mathcal{T} \otimes \mathcal{F}^n)$ относительно сигма-алгебры $\Theta \times \mathcal{F}^n$. Очевидно, $\mathbf{P}(\vartheta \in B) = \mathbf{E}[\mathbf{P}(\vartheta \in B | X^{(n)})]$.

Кроме случая с фиксированным объёмом выборки мы будем рассматривать статистические правила принятия решений, основанные на результатах наблюдений случайной выборки $X^{(\nu)}$ со случайным моментом прекращения наблюдений ν .

Момент прекращения наблюдений ν определяется набором правил $\varphi_s = \{\varphi_n^s(x^{(n)}), x^{(n)} \in \mathbb{X}^n, n \geq 1\}$, задающих вероятность останова на n -ом шаге наблюдений, если к этому моменту был получен результат $x^{(n)} = (x_1, \dots, x_n)$. Здесь $1 - \varphi_n^s(x^{(0)})$ есть вероятность начала проведения экспериментов, $\varphi_n^s(x^{(0)})$ — вероятность принятия решения без проведения наблюдений ($x^{(0)}$ — фиктивное «наблюдение»). В диссертации рассматриваются только правила останова, для которых вероятность останова на каждом шаге равна нулю или единице в зависимости от того, попадает ли выборочный вектор $x^{(n)}$ в область продолжения эксперимента $O_n (\subset \mathbb{X}^n)$ или нет:

$$\varphi_n^s(x^{(n)}) = 1 - \mathbb{I}_{O_n}(x^{(n)}) = \begin{cases} 0, & \text{если } x^{(n)} \in O_n, \\ 1, & \text{если } x^{(n)} \notin O_n, \end{cases}$$

где \mathbb{I}_A — индикаторная функция события A .

Таким образом, момент прекращения наблюдений ν есть измеримое отображение пространства \mathbb{X}^∞ с цилиндрической сигма-алгеброй \mathcal{F}^∞ в пространство натуральных чисел с добавленной бесконечной точкой $\mathbb{N} \cup \{+\infty\}$, определяемое для каждой последовательности $\mathbf{x} = (x_1, x_2, \dots) \in \mathbb{X}^\infty$ следующим образом

$$\nu(\mathbf{x}) = n \Leftrightarrow \{x^{(1)} \in O_1, \dots, x^{(n-1)} \in O_{n-1}, x^{(n)} \notin O_n\}, \quad n \in \mathbb{N}. \quad (1.2)$$

Очевидно, определённая таким образом случайная величина есть марковский момент относительно последовательности сигма-алгебр $\mathcal{F}^n, n = 1, 2, \dots$

Введём следующие переходные вероятности:

$$\begin{aligned} P(\nu = n | \mathbf{x}) &= P(\nu = n | x^{(n)}) = \varphi_n^s(x^{(n)}) \prod_{k=0}^{n-1} (1 - \varphi_n^s(x^{(k)})), \\ P(\nu = \infty | \mathbf{x}) &= 1 - \sum_{n=1}^{\infty} P(\nu = n | x^{(n)}) = \prod_{k=0}^{\infty} (1 - \varphi_n^s(x^{(k)})). \end{aligned} \quad (1.3)$$

Для правил останова, у которых остановка зависит ещё от результата дополнительного (не зависящего от $x^{(n)}$) случайного эксперимента, т.е. в ситуации, когда $\varphi_n^s(x^{(n)})$ может принимать значения, отличные от нуля и единицы, эти формулы дают вероятности останова на шаге с номером $n (\leq \infty)$ для фиксированной последовательности $\mathbf{x} = (x_1, x_2, \dots)$ с учётом возможности

остановки на нулевом шаге. Ясно, что для моментов остановки вида (1.2) эти вероятности равны либо нулю, либо единице.

Условная (при фиксированном θ) вероятность остановки эксперимента на шаге n равна

$$\mathbf{P}_\theta(\nu = n) = \mathbf{E}_\theta[\mathbf{P}(\nu = n | X^{(n)})], \quad n \geq 1,$$

соответственно

$$\mathbf{P}_\theta(\nu = \infty) = 1 - \sum_{n=1}^{\infty} \mathbf{E}_\theta[\mathbf{P}(\nu = n | X^{(n)})].$$

Пусть $\mathbf{h}(\mathbf{x}) = \{h_n : \mathbb{X}^n \mapsto \mathbb{R}_+^1, n \geq 1\}$ — последовательность неотрицательных функций, где при каждом $n \geq 1$ функция h_n измерима относительно σ -алгебры \mathcal{F}^n на \mathbb{X}^n и борелевской σ -алгебры \mathcal{B}_+ на неотрицательной части \mathbb{R}^1 . Если $\mathbf{P}_\theta(\nu = \infty) = 0$, т.е. процедура продолжения наблюдений замкнута (марковский момент ν есть момент остановки), то, используя соотношения (1.3), можно определить условное (при фиксированном θ) математическое ожидание

$$\mathbf{E}_\theta[\mathbf{h}(X^{(\nu)})] = \sum_{n=1}^{\infty} \mathbf{E}_\theta[h_n(X^{(n)}) \mathbf{P}(\nu = n | X^{(n)})]$$

и безусловное математическое ожидание

$$\mathbf{E}[\mathbf{h}(X^{(\nu)})] = \mathbf{E}[\mathbf{E}_\theta[\mathbf{h}(X^{(\nu)})]],$$

где $X^{(\nu)}$ — случайная выборка случайного объёма ν , определяемая как $X^{(\nu)} = X^{(n)}$, если $\nu(X_1, X_2, \dots) = n, n \geq 1$. Аналогично определяется математическое ожидание любой функции (последовательности функций) $\mathbf{h}(\theta, X^{(\nu)})$:

$$\mathbf{E}[\mathbf{h}(\vartheta, X^{(\nu)})] = \sum_{n=1}^{\infty} \mathbf{E}[h_n(\vartheta, X^{(n)}) \mathbf{P}(\nu = n | X^{(n)})]. \quad (1.4)$$

Решение, которое должен принять статистик, состоит в выборе одного из двух альтернативных утверждений $H_0 : \theta \in \Theta_0$ или $H_1 : \theta \in \Theta_1 = \Theta_0^c$. Решение в пользу справедливости гипотезы H_0 будем обозначать через d_0 , а решение в пользу альтернативы через d_1 .

Решение принимается с помощью рандомизированного правила $\varphi_a(\mathbf{x})$ (от английского action) указывающего, с какой вероятностью следует отвергать гипотезу H_0 , если \mathbf{x} — результат наблюдения случайной выборки $X^{(\nu)}$. Точнее,

правило принятия решений φ_a , как и правило остановки, представляет собой последовательность правил: $\varphi_a = \{\varphi_n^a(x^{(n)}), x^{(n)} \in \mathbb{X}^n, n \geq 1\}$.

Правилу φ_a можно сопоставить решающую функция $\delta(\mathbf{x})$ (заданную на пространстве выборочных значений $X^{(\nu)}$ и «рандомизационном» множестве $\mathcal{U} = [0, 1]$), которая при каждом значении \mathbf{x} выбирает решение d_0 или решение d_1 в зависимости от результата u наблюдения вспомогательной случайной величины U с равномерным распределением на отрезке $[0, 1]$: $\delta(\mathbf{x}) = d_1$, если $\varphi_a(\mathbf{x}) \geq u$. Снова, как и выше, δ в действительности есть последовательность $\{\delta_n(x^{(n)}), x^{(n)} \in \mathbb{X}^n, n \geq 1\}$ решающих функций, заданных на выборочных пространствах $\mathbb{X}^n, n \geq 1$.

Пара правил $\langle \varphi_s, \varphi_a \rangle$ (пара статистик $\langle \nu, \delta \rangle$) с моментом остановки ν задаёт статистический критерий. Условную вероятность принятия решения d_0 при фиксированном значении параметра θ (оперативную характеристику) для этого критерия можно вычислить как

$$\begin{aligned} \Psi(d_0 | \theta) &= \mathbf{P}_\theta(\delta(X^{(\nu)}) = d_0) = \mathbf{E}_\theta[1 - \varphi_a(X^{(\nu)})] = \\ &= \sum_{n=1}^{\infty} \mathbf{E}_\theta[(1 - \varphi_n^a(X^{(n)})) \mathbf{P}(\nu = n | X^{(n)})]. \end{aligned}$$

Безусловная вероятность принятия решения d_0 (безусловная оперативная характеристика критерия) вычисляется следующим образом:

$$\Psi(d_0) = \mathbf{P}(\delta(X^{(\nu)}) = d_0) = \mathbf{E}[\mathbf{E}_{\vartheta}(1 - \varphi_a(X^{(\nu)}))] = \mathbf{E}[\Psi(d_0 | \vartheta)].$$

Ясно, что условная и безусловная вероятности принятия решения d_1 равны $\Psi(d_1 | \theta) = 1 - \Psi(d_0 | \theta)$ и $\Psi(d_1) = 1 - \Psi(d_0)$ соответственно.

Пусть $\mathbb{I}_{\Theta_k}(\vartheta)$ — индикаторная функция множества $\Theta_k, k = 0, 1$. Качество критерия $\langle \varphi_s, \varphi_a \rangle$ описывается двумя величинами

$$\begin{aligned} \mathcal{R}_1(\langle \varphi_s, \varphi_a \rangle) &= \mathbf{P}\{\vartheta \in \Theta_0 | \delta = d_1\} = \frac{\mathbf{E}[\mathbb{I}_{\Theta_0}(\vartheta) \varphi_a(X^{(\nu)})]}{\Psi(d_1)}, \\ \mathcal{R}_0(\langle \varphi_s, \varphi_a \rangle) &= \mathbf{P}\{\vartheta \notin \Theta_0 | \delta = d_0\} = \frac{\mathbf{E}[\mathbb{I}_{\Theta_1}(\vartheta) (1 - \varphi_a(X^{(\nu)}))]}{\Psi(d_0)}, \end{aligned}$$

представляющими собой условные вероятности справедливости соответствующих гипотез и называемыми d-апостериорными рисками (коротко, d-рисками) 1-го и 2-го рода. Если решение d_k критерием $\langle \varphi_s, \varphi_a \rangle$ не может быть принято

(безусловная вероятность $\Psi(d_k) = 0$), то значение соответствующего риска полагается по определению равным нулю.

Определение 1.1. Критерий $\langle \varphi_s, \varphi_a \rangle$ называется d-гарантийным, если для заданных ограничений β_0, β_1 выполняются неравенства

$$\mathcal{R}_j(\langle \varphi_s, \varphi_a \rangle) \leq \beta_j, \quad j = 0, 1.$$

Таким образом, основной проблемой d-гарантийного различения гипотез является построение критерия $\langle \varphi_s, \varphi_a \rangle$, который гарантирует заданные ограничения β_0, β_1 на соответствующие d-риски и минимизирует некоторую характеристику момента останова ν , например, математическое ожидание $\mathbf{E}\nu$ или вероятность останова на первых шагах статистического эксперимента и т.п. К сожалению, задача минимизации среднего объёма выборки в классе последовательных процедур при d-гарантийном различении сложных гипотез в общем виде не решена. Задача для простых гипотез может быть сведена к классической задаче А.Вальда (см. [22]). В диссертации проводится сравнение трёх d-гарантийных критериев по некоторым важным характеристикам объёма наблюдений ν .

Замечание 1. Если априорная вероятность какой-либо из проверяемых гипотез достаточно велика:

$$\Pi_0 = \mathbf{P}\{\vartheta \in \Theta_0\} \geq 1 - \beta_0 \quad (\text{или} \quad \Pi_1 = 1 - \Pi_0 \geq 1 - \beta_1),$$

то критерий, принимающий решение без проведения наблюдений в пользу гипотезы $H_0 : \theta \in \Theta_0$ (или $H_1 : \theta \notin \Theta_0$), будет, очевидно, d-гарантийным.

В дальнейшем нам понадобится следующая лемма, в которой d-риски I-го и II-го рода связываются через безусловную вероятность принятия решения d_0 в пользу гипотезы H_0 .

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

$$\mathcal{R}_1(\langle \varphi_s, \varphi_a \rangle) = 1 - \frac{1 - \Pi_0 - \mathcal{R}_0(\langle \varphi_s, \varphi_a \rangle)\Psi(d_0)}{1 - \Psi(d_0)}. \quad (1.5)$$

Доказательство. По определению условной вероятности

$$\begin{aligned} \mathcal{R}_1(\langle \varphi_s, \varphi_a \rangle) &= \mathbf{P}(\vartheta \in \Theta_0 | \boldsymbol{\delta} = d_1) = 1 - \frac{\mathbf{P}(\vartheta \notin \Theta_0, \boldsymbol{\delta} = d_1)}{1 - \Psi(d_0)} = \\ &= 1 - \frac{\mathbf{P}(\vartheta \notin \Theta_0) - \mathbf{P}(\vartheta \notin \Theta_0, \boldsymbol{\delta} = d_0)}{1 - \Psi(d_0)} = \\ &= 1 - \frac{1 - \mathbf{P}(\vartheta \in \Theta_0) - \mathcal{R}_0(\langle \varphi_s, \varphi_a \rangle) \Psi(d_0)}{1 - \Psi(d_0)}. \end{aligned}$$

Лемма доказана. □

1.2 D-гарантийные критерии на фиксированном числе наблюдений

В задаче различения двух гипотез с заданным ограничением на d-риск первого рода для любого правила остановки существует оптимальное правило принятия решений, минимизирующее d-риск второго рода. Это правило было найдено в [21] и его тестовая статистика совпадает с апостериорной вероятностью гипотезы H_0 . Если статистический эксперимент остановлен на шаге с номером n и $\Pi_0(x^{(n)}) = \mathbf{P}(\vartheta \in \Theta_0 | x^{(n)})$ — апостериорная вероятность нулевой гипотезы при условии, что выборка $X^{(n)} = x^{(n)}$, то оптимальное правило принятия решений имеет следующий вид:

$$\varphi_n^a(x^{(n)}) = \begin{cases} 1, & \text{если } \Pi_0(x^{(n)}) < C, \\ \gamma, & \text{если } \Pi_0(x^{(n)}) = C, \\ 0, & \text{если } \Pi_0(x^{(n)}) > C. \end{cases} \quad (1.6)$$

Таким образом, если требуется построить статистическую процедуру с заданными ограничениями на d-риски первого и второго рода, то можно ограничиться рассмотрением только правил принятия решений (1.6).

Если проверка гипотез проводится по данным фиксированного числа наблюдений, то для некоторого $n \geq 1$ правило остановки $\varphi_k^s(x^{(k)}) \equiv 0$ при $k < n$ и $\varphi_k^s(x^{(k)}) \equiv 1$, $k \geq n$. В связи с этим естественно рассмотреть задачу построения гарантийной процедуры тестирования гипотез, обладающей наименьшим объёмом наблюдений.

Минимальный объём наблюдений $n^* = n^*(\beta_0, \beta_1)$, при котором существует d -гарантийный критерий назовём необходимым объёмом выборки (НОВ).

Вывод формул для вычисления НОВ является одной из практически важных задач гарантийного различения гипотез. Следующие секции этого параграфа посвящены решению этой задачи.

Приведём некоторые полезные свойства правила (1.6) (см., например, [68]). Для сокращения записи через $\mathcal{R}_j(n, C, \gamma)$, $j = 0, 1$, будем обозначать d-риск критерия $\langle n, \varphi_a \rangle$ с правилом принятия решения вида (1.6).

Во-первых, заметим, что d-риски этого правила при фиксированном n зависят монотонно от константы C . Точнее, если $C' < C''$ и $\gamma', \gamma'' \in [0, 1]$, то

$$\mathcal{R}_0(n, C', \gamma') \geq \mathcal{R}_0(n, C'', \gamma''), \quad \mathcal{R}_1(n, C', \gamma') \leq \mathcal{R}_1(n, C'', \gamma'').$$

Если при каждом $n \geq 1$ постоянные C_n, γ_n удовлетворяют уравнению

$$\mathcal{R}_1(n, C_n, \gamma_n) = \beta_1, \tag{1.7}$$

то НОВ соответствует тому объёму n , при котором впервые d-риск второго рода решающего правила (1.6) с этими константами будет не больше β_0 :

$$\mathcal{R}_0(n, C_n, \gamma_n) \leq \beta_0. \tag{1.8}$$

Заметим, что равенство (1.7) достижимо только в случае, когда априорная вероятность нулевой гипотезы $\mathbf{P}(\vartheta \in \Theta_0) \geq \beta_1$ и, кроме того, безусловная вероятность $\mathbf{P}(\Pi_0(X^{(n)}) \leq \beta_1) > 0$.

Рассмотренные далее вероятностные модели обладают монотонным отношением правдоподобия. В этом случае при проверке односторонних гипотез, скажем, $H_0 : \theta \leq \theta_0$, $H_1 : \theta > \theta_0$ о параметре $\theta \in \mathbb{R}^1$ оптимальное правило принятия решения может быть представлено через соответствующую статистику.

Лемма 1.2. *Если вероятностная модель наблюдений $f_n(x^{(n)} | \theta)$, $\theta \in \Theta \subset \mathbb{R}^1$, имеет монотонно возрастающее отношение правдоподобия относительно статистики $T = T(X^{(n)})$, то при различении гипотез $H_0 : \theta \leq \theta_0$, $H_1 : \theta > \theta_0$ оптимальное правило (1.6) эквивалентно правилу*

$$\varphi_a(x^{(n)}) = \begin{cases} 1, & \text{если } T(x^{(n)}) > t_0, \\ \gamma_0, & \text{если } T(x^{(n)}) = t_0, \\ 0, & \text{если } T(x^{(n)}) < t_0 \end{cases} \tag{1.9}$$

с некоторыми константами γ_0, t_0 .

Замечание 2. Если отношение правдоподобия есть убывающая функция T , то оптимальное правило рекомендует с вероятностью единица принимать гипотезу $H_0 : \theta \leq \theta_0$ при $T(x^{(n)}) > t_0$.

Проиллюстрируем предложенные методы построения d-гарантийных критериев с необходимым объёмом выборки для ряда конкретных вероятностных моделей, обычно используемых при проведении контроля качества.

Нормально-нормальная модель эксперимента. Рассмотрим задачу различения двух односторонних гипотез $H_0 : \theta \leq \theta_0$ и $H_1 : \theta > \theta_0$ о среднем значении $\theta \in \mathbb{R}^1$ нормального $\mathcal{N}(\theta, \sigma^2)$ распределения с известной дисперсией σ^2 . Параметр θ есть реализация случайной величины ϑ с априорным нормальным $\mathcal{N}(\mu, \tau^2)$ распределением. Символически эту вероятностную модель, на которую в дальнейшем будем ссылаться как на модель N–N, можно представить в виде

$$\vartheta \sim \mathcal{N}(\mu, \tau^2), \quad X \underset{\vartheta=\theta}{\sim} \mathcal{N}(\theta, \sigma^2). \quad (1.10)$$

Обозначим через Φ — функцию распределения стандартного нормального $(0, 1)$ закона, а через ϕ — её функцию плотности.

Параметры μ, τ^2, σ^2 будем считать известными. В рамках эмпирического d-апостериорного подхода последнее означает, что эти параметры могут быть оценены по имеющемуся в наличии архиву предыдущих наблюдений (см. далее раздел 1.4.1).

Поскольку θ есть параметр сдвига, то, не ограничивая общности, можно положить $\theta_0 = 0$. Таким образом, априорная вероятность справедливости нулевой гипотезы $\Pi_0 = \Phi(-\mu/\tau) = 1 - \Phi(\mu/\tau)$.

Данная модель обладает монотонным отношением правдоподобия относительно выборочного среднего $T = \bar{X}$. Хорошо известно (см., например, [17]), что при фиксированном $\bar{X} = \bar{x}$ условное распределение ϑ нормально $\mathcal{N}(m_n, \tau_n^2)$, где

$$m_n = m_n(\bar{x}) = \frac{\mu \sigma^2 + n \bar{x} \tau^2}{\sigma^2 + n \tau^2}, \quad \tau_n^2 = \frac{\sigma^2 \tau^2}{\sigma^2 + n \tau^2}. \quad (1.11)$$

Безусловное распределение \bar{X} (как свёртка двух нормальных законов) также нормально: $\bar{X} \sim \mathcal{N}(\mu, \tau^2 + \sigma^2/n)$, т.е. для любого $t \in \mathbb{R}^1$ безусловная

вероятность $\mathbf{P}(\bar{X} < t) = \Phi((t - \mu)/\sqrt{\tau^2 + \sigma^2/n})$.

В силу непрерывности распределения статистики \bar{X} d-риск оптимального правила (1.9) не зависит от константы рандомизации γ_0 . Следовательно, можно считать $\gamma_0 = 1$, т.е. правило принятия решений (1.9), соответствующее НОВ, будет принимать нулевую гипотезу $H_0 : \theta \leq 0$, если выборочное среднее $\bar{X} < t$. Величины d-рисков 1-го и 2-го рода процедуры с таким правилом обозначим как

$$\mathcal{R}_1(n, t) := \mathbf{P}(\vartheta \leq \theta_0 | \bar{X} \geq t), \quad \mathcal{R}_0(n, t) := \mathbf{P}(\vartheta > \theta_0 | \bar{X} < t)$$

соответственно.

Константа t выбирается по заданному ограничению на величину d-риска. Из описанных выше свойств следует, что d-риск $\mathcal{R}_0(n, t)$ есть непрерывная строго возрастающая функция $t \in \mathbb{R}^1$, причём

$$\lim_{t \rightarrow -\infty} \mathcal{R}_0(n, t) = 0, \quad \lim_{t \rightarrow +\infty} \mathcal{R}_0(n, t) = 1 - \Pi_0.$$

Следовательно, если априорная вероятность справедливости нулевой гипотезы $\Pi_0 = 1 - \Phi(\mu/\tau) < 1 - \beta_0$, то для каждого $n \geq 1$ существует единственная константа $t = t_n$, удовлетворяющая уравнению $\mathcal{R}_0(n, t_n) = \beta_0$, которое в данном случае можно записать в виде

$$\begin{aligned} \mathcal{R}_0(n, t) &= \\ &= \frac{1}{\Phi((t - \mu)/\sqrt{\tau^2 + \sigma^2/n})} \int_0^\infty \Phi\left(\frac{t - \theta}{\sigma/\sqrt{n}}\right) \frac{1}{\tau} \phi\left(\frac{\theta - \mu}{\tau}\right) d\theta = \beta_0. \end{aligned} \quad (1.12)$$

Необходимый объём выборки n^* соответствует тому n , при котором впервые будет иметь место неравенство $\mathcal{R}_1(n, t_n) \leq \beta_1$ или (в эквивалентной форме с $t = t_n$)

$$\begin{aligned} \mathcal{R}_1(n, t) &= \\ &= \frac{1}{\Phi((\mu - t)/\sqrt{\tau^2 + \sigma^2/n})} \int_{-\infty}^0 \Phi\left(\frac{\theta - t}{\sigma/\sqrt{n}}\right) \frac{1}{\tau} \phi\left(\frac{\theta - \mu}{\tau}\right) d\theta \leq \beta_1. \end{aligned} \quad (1.13)$$

Апостериорная вероятность нулевой гипотезы $\Pi_0(x^{(n)}) = \Phi(-m_n(\bar{x})/\tau_n)$, поэтому правило, принимающее решение d_0 , когда $\Pi_0(x^{(n)}) > \Phi(-c_0)$ с некоторым c_0 , будет принимать нулевую гипотезу, если

$$\bar{x} < -\mu \frac{\sigma^2}{n\tau^2} + c_0 \sqrt{1 + \frac{\sigma^2}{n\tau^2}} \frac{\sigma}{\sqrt{n}}. \quad (1.14)$$

В случае, когда $\Phi(\mu/\tau) \leq \beta_0$ (или $\Phi(\mu/\tau) \geq 1 - \beta_1$), гипотеза H_0 (соответственно H_1) может быть принята без проведения наблюдений.

Замечание 3. Из формул (1.12)–(1.13) понятно, что между НОВ n^* и дисперсией наблюдений σ^2 существует простая связь $n^* = const \cdot \sigma^2$ (с поправкой на дискретность n^*). Поэтому поведение n^* достаточно изучать только при $\sigma^2 = 1$. Кроме того, теоретический анализ удобнее проводить также при $\tau^2 = 1$. Легко понять, что при формулировке результатов в общем случае достаточно заменить μ на μ/τ , объём выборки n на $n\sigma^2/\tau^2$, а область принятия нулевой гипотезы вида $\bar{X} < t$ на $\bar{X} < \tau t$. Если $\mu = 0$ ($= \theta_0$), то НОВ n^* обратно пропорционален τ^2 (см. ещё (1.15) ниже). Если же $\mu \neq 0$, то при достаточно малой дисперсии τ^2 априорная вероятность $\Pi_0 \geq 1 - \beta_0$ (или $\Pi_0 \leq \beta_1$) и НОВ $n^* = 0$.

Для модели N–N в одном важном частном случае можно привести точную формулу для НОВ n^* . Пусть $[a]$ — целая часть положительного числа a с округлением вверх, т.е. наименьшее целое число, не меньше a .

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

$$n^* = \left\lceil \frac{\sigma^2}{\tau^2 \operatorname{tg}^2(\pi\beta)} \right\rceil. \quad (1.15)$$

Доказательство. Левые части соотношений (1.12)–(1.13) представляют собой непрерывные функции параметров t и n , если формально n считать непрерывной переменной. Таким образом, в данном случае НОВ $n^* = [n]$, если действительное число n выбрать как решение системы (1.12)–(1.13), в которой знак неравенства „ \leq ” заменён на знак равенства. Выберем t и n как решения системы уравнений (1.12)–(1.13). Тогда в соответствии с (1.5), где для случая $\mu = \theta_0$ априорная вероятность $\Pi_0 = 1/2$, получим равенство

$$\beta = 1 - \frac{1/2 - \beta \Psi(d_0)}{1 - \Psi(d_0)},$$

с $\Psi(d_0) = \Phi(-t/\sqrt{\tau^2 + \sigma^2/n})$. Это равенство возможно только при $\Psi(d_0) = 1/2$, т.е. параметр $t = 0$.

Таким образом, «непрерывный» параметр n должен удовлетворять уравнению (1.12), которое в данном случае приобретает вид

$$\int_0^{\infty} \Phi(-\theta\sqrt{n}/\sigma) \frac{1}{\tau} \phi(\theta/\tau) d\theta = \frac{\beta}{2}.$$

Интеграл в левой части здесь можно записать в виде двойного интеграла

$$\frac{1}{2\pi} \int_0^{\infty} \int_{-\infty}^{-r\theta} \exp\left(-\frac{1}{2}(\theta^2 + z^2)\right) dz d\theta,$$

где $r = \tau\sqrt{n}/\sigma$. После стандартной сферической замены переменных получаем, что этот интеграл равен $\arctan(1/r)/2\pi$. Отсюда уже легко получается равенство (1.15). \square

Следующее утверждение упрощает отыскание НОВ n^* и соответствующей ему критической константы $t = t(n^*)$. Справедлива

Лемма 1.3. *Для модели N–N при проверке гипотезы $H_0 : \theta \leq 0$ против $H_1 : \theta > 0$ с $\beta_0 < \Phi(\mu/\tau) < 1 - \beta_1$ необходимый объём выборки n^* совпадает с целой частью $\lceil z \rceil$ решения уравнения*

$$\int_0^{\infty} \Phi\left(\frac{t(z) - \theta}{\sigma/\sqrt{z}}\right) \frac{1}{\tau} \phi\left(\frac{\theta - \mu}{\tau}\right) d\theta = \beta_0 \frac{\Pi_0 - \beta_1}{1 - \beta_0 - \beta_1}, \quad (1.16)$$

$$t(z) = \mu + \sqrt{\tau^2 + \frac{\sigma^2}{z}} \Phi^{-1}\left(\frac{\Pi_0 - \beta_1}{1 - \beta_0 - \beta_1}\right), \quad z > 0. \quad (1.17)$$

Доказательство. Аналогично доказательству предыдущей теоремы будем считать n непрерывной переменной, то есть НОВ n^* есть целая часть решения системы (1.12)–(1.13) со знаком равенства в обоих соотношениях. Так как здесь достигаются равенства ограничений на риски процедуры, то в силу тождества (1.5) вероятность принятия решения d_0

$$\Psi(d_0) = \Phi\left(\frac{t - \mu}{\sqrt{\tau^2 + \sigma^2/n}}\right) = \frac{\Pi_0 - \beta_1}{1 - \beta_0 - \beta_1},$$

что эквивалентно (1.17). С учётом этого равенства уравнение (1.16) эквивалентно (1.12). Снова применяя тождество (1.5), получаем, что в этом случае будет выполняться и равенство в (1.13). \square

Наименее информативное априори. В байесовском анализе, когда нет оснований рассматривать априорное распределение параметра какого-либо конкретного типа, принято выбирать наименее информативное распределение. В качестве последнего часто берётся равномерное распределение (мера Лебега, см., например, [27]). Понятно, что на \mathbb{R}^1 равномерное априорное «распределение» не позволит вычислить значение d-риска. Поэтому, как некое приближение к равномерности, можно рассмотреть схему с нормальным априорным распределением и дисперсией $\tau^2 \rightarrow \infty$. К сожалению, для d-апостериорного подхода равномерное распределение скорее наиболее, чем наименее, информативное. Справедлива

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

Доказательство. Во-первых, заметим, что при $\tau^2 \rightarrow \infty$ априорная вероятность нулевой гипотезы $\Pi_0 = \Phi(-\mu/\tau) \rightarrow 1/2$. Следовательно, НОВ $n^* > 0$. Далее, т.к. параметр σ^2 не играет существенной роли, то в дальнейших выкладках, положим $\sigma^2 = 1$. Произведем замену $\theta \rightarrow \tau z$, $z \in [0, \infty)$, переменной интегрирования в уравнении (1.12):

$$\frac{1}{\Phi((\tilde{t} - \mu/\tau)/\sqrt{1 + 1/n\tau^2})} \int_0^\infty \Phi((\tilde{t} - z)\sqrt{n}\tau)\phi(z - \mu/\tau) dz = \beta_0,$$

где $\tilde{t} = t/\tau$. Зафиксируем \tilde{t} и положим $\tau^2 \rightarrow \infty$. Поскольку подынтегральное выражение в предыдущем уравнении при достаточно малом значении μ/τ ограничено интегрируемой функцией, скажем $2\phi(z/2)$, то по теореме Лебега о мажорируемой сходимости здесь можно перейти к пределу под знаком интеграла. Так как $\Phi((\tilde{t} - z)\sqrt{n}\tau)\phi(z - \mu/\tau) \rightarrow \mathbb{I}(z < \tilde{t})\phi(z)$, то получаем следующее уравнение (относительно \tilde{t})

$$\frac{\Phi(\tilde{t}) - 1/2}{\Phi(\tilde{t})} = \beta_0,$$

т.е. $\tilde{t} = \Phi^{-1}(1/(2(1 - \beta_0))) > 0$ при $\beta_0 < 1/2$.

Поскольку $\tilde{t} > 0$, то $\Phi((z - \tilde{t})\sqrt{n}\tau)\phi(z - \mu/\tau) \rightarrow 0$ для любых $z < 0$. Следовательно, для любого фиксированного $n \geq 1$ при $\tau^2 \rightarrow \infty$ левая часть

неравенства (1.13) стремится к нулю. Другими словами, любое $n \geq 1$ при достаточно большом $\tau^2 (\geq \tau_0^2)$ становится достаточным объёмом выборки. Что и требовалось доказать. \square

Показательная модель наблюдений с гамма-распределением параметра интенсивности отказов. Рассмотрим теперь задачу различения двух односторонних гипотез $H_0 : \theta \leq \theta_0$, $H_1 : \theta > \theta_0$ по выборке $X^{(n)}$ фиксированного объёма n из показательного распределения $\mathcal{E}(\theta)$ с плотностью $f(x | \theta) = \theta \exp\{-\theta x\}$, $x > 0$. Параметр θ есть реализация случайной величины ϑ с априорным гамма-распределением $\mathcal{G}(\lambda, a)$ с плотностью

$$g(\theta; \lambda, a) = \frac{a^\lambda}{\Gamma(\lambda)} \theta^{\lambda-1} e^{-a\theta}, \quad \theta > 0.$$

Пусть $\mathbb{G}(\cdot; \lambda, a)$ — функция распределения соответствующего гамма-закона. Будем обозначать данную модель как G–E. Символически эту модель можно записать как

$$\vartheta \sim \mathcal{G}(\lambda, a), \quad X \underset{\vartheta=\theta}{\sim} \mathcal{E}(\theta). \quad (1.18)$$

Показательная вероятностная модель обладает монотонным отношением правдоподобия относительно статистики $1/S_n$, где $S_n = \sum_1^n X_i$. Ясно, что все построения можно осуществлять, опираясь на статистику S_n . Заметим, что безусловное распределение S_n может быть представлено через функцию распределения бета-закона с параметрами (n, λ) (будем обозначать эту функцию через $\mathbb{B}(\cdot; n, \lambda)$). Такое представление значительно облегчает вычисление интегралов, содержащихся в выражениях для d-рисков.

Лемма 1.4. *Если $\vartheta \sim \mathcal{G}(\lambda, a)$, $X \underset{\vartheta=\theta}{\sim} \mathcal{E}(\theta)$, то*

а) безусловная функция распределения S_n

$$\mathbf{P}(S_n < t) = \mathbb{B}\left(\frac{t}{t+a}; n, \lambda\right) = \frac{\Gamma(n+\lambda)}{\Gamma(n)\Gamma(\lambda)} \int_0^{t/(t+a)} x^{n-1} (1-x)^{\lambda-1} dx;$$

б) апостериорное распределение ϑ при фиксированном $S_n = t$ есть гамма-распределение $\mathcal{G}(n+\lambda, t+a)$.

Доказательство. а) Так как при фиксированном θ статистика $S_n \sim \mathcal{G}(n, \theta)$, то безусловная (частная) функция распределения

$$\begin{aligned} \mathbf{P}(S_n < t) &= \mathbf{E}[\mathbf{P}(S_n < t | \vartheta)] = \\ &= \frac{a^\lambda}{\Gamma(n)\Gamma(\lambda)} \int_0^\infty d\theta \int_0^t \theta^n x^{n-1} e^{-\theta x} \theta^{\lambda-1} e^{-a\theta} dx = \\ &= \frac{a^\lambda}{\Gamma(n)\Gamma(\lambda)} \int_0^t \left[\int_0^\infty \theta^{n+\lambda-1} e^{-\theta(x+a)} d\theta \right] x^{n-1} dx = \\ &= \frac{a^\lambda \Gamma(n+\lambda)}{\Gamma(n)\Gamma(\lambda)} \int_0^t \frac{x^{n-1}}{(x+a)^{n+\lambda}} dx = \frac{\Gamma(n+\lambda)}{\Gamma(n)\Gamma(\lambda)} \int_0^{t/(t+a)} u^{n-1} (1-u)^{\lambda-1} du \end{aligned}$$

после замены $x/(x+a) \rightarrow u$.

б) Плотность совместного распределения (ϑ, S_n) (относительно переменных (θ, t)) равна

$$\frac{a^\lambda t^{n-1}}{\Gamma(n)\Gamma(\lambda)} \theta^{n+\lambda-1} e^{-\theta(t+a)}, \quad \theta, t \geq 0.$$

Следовательно, апостериорная плотность ϑ пропорциональна $\theta^{n+\lambda-1} e^{-\theta(t+a)}$, что доказывает утверждение леммы. \square

Критерий, соответствующий НОВ, отвергает гипотезу $H_0 : \theta \leq \theta_0$, если сумма выборочных значений $S_n \leq t$. Совместная вероятность

$$\begin{aligned} \mathbf{P}(\vartheta \leq \theta_0, S_n \leq t) &= \int_0^{\theta_0} \mathbb{G}(t\theta; n, 1) g(\theta; \lambda, a) d\theta = \\ &= \int_0^{a\theta_0} \mathbb{G}(t\theta/a; n, 1) g(\theta; \lambda, 1) d\theta. \end{aligned}$$

Константа $t = t_n$ при каждом фиксированном объёме наблюдений n ($= 1, 2, \dots$) определяется из условия на d-риск второго рода

$$\mathcal{R}_1(n, t) = \frac{1}{\mathbb{B}\left(\frac{t}{t+a}; n, \lambda\right)} \int_0^{a\theta_0} \mathbb{G}(t\theta/a; n, 1) g(\theta; \lambda, 1) d\theta = \beta_1.$$

Необходимый объём выборки n^* находится как наименьшее n , удовлетворяющее соответствующему неравенству на d-риск первого рода:

$$\mathcal{R}_0(n, t) := \frac{1}{1 - \mathbb{B}\left(\frac{t}{t+a}; n, \lambda\right)} \int_{a\theta_0}^\infty (1 - \mathbb{G}(t\theta/a; n, 1)) g(\theta; \lambda, 1) d\theta \leq \beta_0.$$

Замечание 4. Из представленных формул видно, что можно ограничиться рассмотрением гипотез с граничной точкой $\theta_0 = 1$ — общий случай получается заменой $a \dashrightarrow a\theta_0$, $t \dashrightarrow t\theta_0$.

Как и в нормальном случае, здесь также задачу отыскания n^* можно свести к задаче решения системы из двух уравнений $\mathcal{R}_0(n, t) = \beta_0$, $\mathcal{R}_1(n, t) = \beta_1$ относительно непрерывных параметров n, t . В этом случае, в силу тождества (1.5) вероятность принятия гипотезы H_1

$$\Psi(d_1) = \mathbb{B}\left(\frac{t}{t+a}; n, \lambda\right) = \frac{1 - \Pi_0 - \beta_0}{1 - \beta_0 - \beta_1},$$

где $\Pi_0 = \mathbb{G}(\theta_0; \lambda, a) = \mathbb{G}(a; \lambda, 1)$ — вероятность справедливости нулевой гипотезы (с $\theta_0 = 1$). Таким образом, задача поиска НОВ сводится к решению относительно (n, t) системы уравнений

$$\int_0^a \mathbb{G}(t\theta; n, 1) g(\theta; \lambda, 1) d\theta = \beta_1 \frac{1 - \Pi_0 - \beta_0}{1 - \beta_0 - \beta_1}, \quad (1.19)$$

$$\mathbb{B}\left(\frac{t}{t+a}; n, \lambda\right) = \frac{1 - \Pi_0 - \beta_0}{1 - \beta_0 - \beta_1}. \quad (1.20)$$

При построении последовательных и асимптотических процедур для этой модели нам понадобится асимптотическое представление 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 $\mathcal{G}(n, 1)$ имеет асимптотическое (при $n \rightarrow \infty$) представление

$$\mathbb{G}^{-1}(\gamma; n, 1) = n + \sqrt{nt_\gamma} + \frac{1}{3}(t_\gamma^2 - 1) + o(1). \quad (1.21)$$

Доказательство. Гамма-закон $\mathcal{G}(n, 1)$ с целым показателем n описывает распределение суммы n независимых показательных величин ξ_1, \dots, ξ_n с единичным параметром масштаба. Поэтому для функции распределения этого закона можно применить разложение в ряд Эджворта (см. [24, Гл. XVI, § 4])

$$\begin{aligned} \mathbb{G}(n + x\sqrt{n}; n, 1) &= \mathbf{P}\left(\sum_{i=1}^n (\xi_i - 1) < x\sqrt{n}\right) = \\ &= \Phi(x) + \phi(x) \frac{\alpha_3(1 - x^2)}{6} \frac{1}{\sqrt{n}} + C_n(x) \frac{1}{\sqrt{n}}, \end{aligned} \quad (1.22)$$

где $\sup_x |C_n(x)| \xrightarrow{n} 0$; третий момент показательной случайной величины $\alpha_3 = \mathbf{E}(\xi_1 - 1)^3 = 2$.

Пусть x_n^* , x_n , t_γ таковы, что

$$\begin{aligned}\Phi(t_\gamma) &= \gamma, \\ \mathbb{G}(n + x_n^* \sqrt{n}; n, 1) &= \gamma, \\ \Phi(x_n) + \frac{1}{\sqrt{n}} P(x_n) &= \gamma\end{aligned}$$

с $P(x) = \phi(x)(1 - x^2)/3$. Очевидно, $x_n \rightarrow t_\gamma$, а в силу (1.22) и $x_n^* \rightarrow t_\gamma$ при $n \rightarrow \infty$. Кроме того, по формуле конечных приращений из (1.22) имеем

$$\begin{aligned}C_n(x_n^*) &= \left(\Phi(x_n) - \Phi(x_n^*) + (P(x_n) - P(x_n^*)) \frac{1}{\sqrt{n}} \right) \sqrt{n} = \\ &= \sqrt{n}(x_n - x_n^*) \left[\Phi'(\tilde{x}_n) + P'(\hat{x}_n) \frac{1}{\sqrt{n}} \right],\end{aligned}$$

где \tilde{x}_n , $\hat{x}_n \rightarrow t_\gamma$. Так как выражение в квадратных скобках при $n \rightarrow \infty$ сходится к $\Phi'(t_\gamma) = \phi(t_\gamma) \neq 0$ и $C_n(x_n^*) \rightarrow 0$, то $\sqrt{n}(x_n - x_n^*) \rightarrow 0$.

Аналогично, по формуле конечных приращений

$$0 = \Phi'(\tilde{x}_n)(x_n - t_\gamma) + P(x_n) \frac{1}{\sqrt{n}},$$

где $\tilde{x}_n \rightarrow t_\gamma$. Таким образом, при $n \rightarrow \infty$

$$\sqrt{n}(x_n - t_\gamma) = -\frac{\phi(x_n)(1 - x_n^2)/3}{\Phi'(\tilde{x}_n)} = \frac{t_\gamma^2 - 1}{3} + o(1),$$

т.е. $x_n = t_\gamma + (t_\gamma^2 - 1)/(3\sqrt{n}) + o(1/\sqrt{n})$. По построению $\mathbb{G}^{-1}(\gamma; n, 1) = n + x_n^* \sqrt{n} = n + x_n \sqrt{n} + o(1) = n + \sqrt{n} t_\gamma + \frac{1}{3}(t_\gamma^2 - 1) + o(1)$, что и требовалось. \square

Замечание 5. Для нецелых значений параметра формы гамма-распределения эта формула также справедлива. Численные расчёты показывают, что при $0.001 \leq \gamma \leq 0.999$ и $n = 1$ ошибка приближения не превышает 0.76 и эта ошибка убывает с ростом n пропорционально $1/\sqrt{n}$.

Испытания Бернулли с априорным бета-распределением. В статистическом эксперименте наблюдается бернуллиевская случайная величина X , принимающая значения 1 и 0 с вероятностями θ и $1 - \theta$ соответственно. В данном случае в качестве априорного распределения вполне оправдано

выбрать бета-распределение, поскольку богатство форм этой модели обеспечивает хорошую подгонку реального априорного распределения. Итак, априорное распределение имеет функцию плотности

$$g(\theta; p, q) = \frac{1}{\mathbb{B}(p, q)} \theta^{p-1} (1 - \theta)^{q-1}, \quad 0 < \theta < 1,$$

где $\mathbb{B}(p, q) = \Gamma(p)\Gamma(q)/\Gamma(p + q)$ — бета-функция Эйлера. Как и выше, через $\mathbb{B}(t; a, b)$, $t \in [0, 1]$, будем обозначать функцию распределения бета-закона с параметрами (a, b) . Символически эту модель с аббревиатурой В-В будем записывать как

$$\vartheta \sim \mathcal{B}(p, q), \quad X \underset{\vartheta=\theta}{\sim} \mathcal{Bern}(\theta). \quad (1.23)$$

Рассматривается задача различения односторонних гипотез $H_0 : \theta \leq \theta_0$, $H_1 : \theta > \theta_0$. Модель Бернулли обладает монотонным отношением правдоподобия относительно статистики $S_n = \sum_1^n X_i$. Ясно, что оптимальный критерий должен быть рандомизированным. Этот критерий будет принимать нулевую гипотезу с вероятностью 1, если $S_n < t$ для некоторой константы t , и с вероятностью $\gamma \in [0, 1]$, если $S_n = t$. При $S_n > t$ нулевая гипотеза отвергается с вероятностью 1.

Функция безусловного распределения S_n равна

$$\begin{aligned} F_n(t; p, q) &:= \mathbf{P}\{S_n < t\} = \\ &= \sum_{k=0}^{t-1} \int_0^1 C_n^k \theta^k (1 - t)^{n-k} \frac{1}{\mathbb{B}(p, q)} \theta^{p-1} (1 - \theta)^{q-1} d\theta = \sum_{k=0}^{t-1} C_n^k \omega(p, q, n, k), \end{aligned} \quad (1.24)$$

где

$$\omega(p, q, n, k) = \frac{\mathbb{B}(p + k, q + n - k)}{\mathbb{B}(p, q)}.$$

Следовательно, $\mathbf{P}(S_n = k) = \omega(p, q, n, k) C_n^k$, $k = 0, 1, \dots, n$.

Аналогично записывается совместная вероятность

$$\begin{aligned} \mathbf{P}\{\vartheta > \theta_0, S_n < t\} &= \sum_{k=0}^{t-1} \int_{\theta_0}^1 C_n^k \theta^k (1 - \theta)^{n-k} \frac{1}{\mathbb{B}(p, q)} \theta^{p-1} (1 - \theta)^{q-1} d\theta = \\ &= \sum_{k=0}^{t-1} C_n^k \omega(p, q, n, k) (1 - \mathbb{B}(\theta_0; p + k, q + n - k)). \end{aligned}$$

Апостериорная вероятность справедливости нулевой гипотезы при фиксированном значении $S_n = t$ вычисляется как

$$\mathbf{P}(\vartheta \leq \theta_0 | S_n = t) = \mathbb{B}(\theta_0; p + t, q + n - t). \quad (1.25)$$

Рандомизационную составляющую процесса принятия решения можно включить в вероятностную модель следующим образом. Обозначим через $[c]$ целую часть числа c , через $\{c\}$ — его дробную часть. Пусть случайная величина U имеет равномерное распределение на отрезке $[0, 1]$ и не зависит от ϑ и X . Критерий будет принимать гипотезу $H_0 : \theta \leq \theta_0$, если $S_n + U < t$, и отвергать, если $S_n + U \geq t$. Другими словами, если $S_n < [t]$, то и $S_n + U < [t]$ при любом значении U , т.е. при $S_n < [t]$ с вероятностью единица нулевая гипотеза принимается. Если же $S_n = [t]$, то с вероятностью $\{t\} = t - [t]$ величина U примет значение, при котором $S_n + U < t$, т.е. гипотеза H_0 будет принята с вероятностью $\{t\}$. Если t — целое число, то с вероятностью 1 гипотеза принимается, когда $S_n < t$; в противном случае гипотеза отвергается.

Совместная вероятность

$$\begin{aligned} \mathbf{P}(\vartheta > \theta_0, S_n + U < t) &= \mathbf{P}(\vartheta > \theta_0, S_n < [t]) + \{t\}\mathbf{P}(\vartheta > \theta_0, S_n = [t]) = \\ &= \sum_{k=0}^{[t]-1} C_n^k \omega(p, q, n, k) (1 - \mathbb{B}(\theta_0; p + k, q + n - k)) + \\ &\quad + \{t\} C_n^{[t]} \omega(p, q, n, [t]) (1 - \mathbb{B}(\theta_0; p + [t], q + n - [t])). \end{aligned}$$

Безусловная вероятность принятия гипотезы H_0 вычисляется как

$$\begin{aligned} \mathbf{P}(S_n + U < t) &= \mathbf{P}(S_n < [t]) + \{t\}\mathbf{P}(S_n = [t]) = \\ &= F_n([t]; p, q) + \{t\} C_n^{[t]} \omega(p, q, n, [t]). \end{aligned}$$

Константа t может принимать значения из отрезка $[0, n + 1]$ и выбирается при каждом фиксированном объёме наблюдений $n (= 1, 2, \dots)$ из условия на величину d-риска второго рода:

$$\mathcal{R}_0(n, t) = \frac{\mathbf{P}(\vartheta > \theta_0, S_n + U < t)}{\mathbf{P}(S_n + U < t)} = \beta_0.$$

Необходимый объём выборки n^* определяется как наименьшее n , удовлетворяющее неравенству

$$\mathcal{R}_1(n, t) = \frac{\mathbf{P}(\vartheta \leq \theta_0, S_n + U \geq t)}{1 - \mathbf{P}(S_n + U < t)} \leq \beta_1,$$

где

$$\begin{aligned}
\mathbf{P}(\vartheta \leq \theta_0, S_n + U \geq t) &= \\
&= \mathbf{P}(\vartheta \leq \theta_0, S_n > [t]) + (1 - \{t\})\mathbf{P}(\vartheta \leq \theta_0, S_n = [t]) = \\
&= \sum_{k=[t]+1}^n C_n^k \mathbb{B}(\theta_0; p+k, q+n-k) \omega(p, q, n, k) + \\
&\quad + (1 - \{t\})C_n^{[t]} \mathbb{B}(\theta_0; p+[t], q+n-[t]) \omega(p, q, n, [t]).
\end{aligned}$$

Замечание 6. Рандомизация для этой дискретной модели наблюдений может сильно влиять на величину НОВ. Пусть $\theta_0 = 0.01$ и параметры априорного распределения $p = 0.1$, $q = 10$. Необходимый объём выборки при $\beta_0 = \beta_1 = 0.049$ для критерия с рандомизацией равен $n^* = 165$; критическая константа $t = 3.162$, т.е. партия принимается, если среди контрольных испытаний либо меньше 3 бракованных изделий, либо 3, но в этом случае партия принимается только с вероятностью 0.162. Если же отказаться от возможности рандомизации, то придётся проводить уже 211 испытаний — партия принимается, если в ней не больше 3 бракованных изделий.

1.2.1 Асимптотика необходимого объёма выборки в схеме с жёсткими ограничениями на d-риски

В проблеме различения двух односторонних гипотез $H_0 : \theta \leq \theta_0$, $H_1 : \theta > \theta_0$ о параметре $\theta \in \mathbb{R}^1$ существуют два асимптотических подхода к аппроксимации НОВ. В обоих подходах асимптотика НОВ обратно пропорциональна информационному количеству Фишера в точке θ_0

$$I_0 = \mathbf{E}_\theta \left(\frac{\partial \ln f(X | \theta)}{\partial \theta} \right)^2 \Big|_{\theta=\theta_0}.$$

В первом подходе асимптотическая формула даёт приближённое значение НОВ при $\beta_0, \beta_1 \rightarrow 0$. В [58, теорема 2.2] показано, что при условиях регулярности к вероятностной модели необходимый объём выборки

$$n^* \sim n_\beta = \begin{cases} \frac{g^2(\theta_0)}{I_0} \frac{c_0^2}{(\beta_1 \Pi_1 - \beta_0 \Pi_0)^2}, & \text{если } \Pi_1 \beta_1 \neq \Pi_0 \beta_0, \\ \frac{g^2(\theta_0)}{2\pi I_0} \frac{1}{(\Pi_0 \beta_0)^2}, & \text{если } \Pi_1 \beta_1 = \Pi_0 \beta_0, \end{cases} \quad (1.26)$$

где константа c_0 есть решение уравнения

$$\frac{\beta_1 \Pi_1}{\beta_0 \Pi_0} = \frac{W(-c_0)}{W(c_0)}, \quad (1.27)$$

$W(c) = \phi(c) + c\Phi(c)$, $c \in (-\infty, \infty)$. Критическая константа C для правила (1.6) также имеет конечное предельное значение:

$$\lim_{\beta_0 + \beta_1 \rightarrow 0} C = C_0 = \Phi(-c_0).$$

Для вывода уточнения асимптотической формулы (1.26) нам понадобится другое представление константы c_0 .

Лемма 1.6. Уравнение (1.27) эквивалентно уравнению

$$\frac{\Pi_0}{(1 + \rho)\Pi_0 - \rho} c_0 = W(c_0), \quad (1.28)$$

где $\rho = \beta_1/\beta_0$.

Доказательство. В силу симметрии стандартной нормальной функции распределения Φ и чётности ее функции плотности ϕ справедливо $W(-c) = \phi(c) - c(1 - \Phi(c)) = W(c) - c$. Следовательно, уравнение (1.27) можно переписать как $\rho\Pi_1/\Pi_0 = 1 - c_0/W(c_0)$, что эквивалентно (1.28). \square

Выражение для n_β также можно представить в более компактной форме.

Лемма 1.7. Пусть c_0 — решение уравнения (1.27), тогда

$$n_\beta = \frac{W^2(c_0) g^2(\theta_0)}{I_0} \frac{1}{(\Pi_0 \beta_0)^2} = \frac{W^2(-c_0) g^2(\theta_0)}{I_0} \frac{1}{(\Pi_1 \beta_1)^2}. \quad (1.29)$$

Доказательство. Второе равенство (1.29) есть следствие (1.27). Далее, легко показать (с помощью интегрирования по частям), что

$$W(c) = \int_{-\infty}^c \Phi(u) du. \quad (1.30)$$

Отсюда понятно, что функция W есть строго возрастающая непрерывная функция $c \in \mathbb{R}^1$. При этом $W(-\infty) = 0$, $W(+\infty) = +\infty$. Следовательно, в случае $\Pi_1 \beta_1 = \Pi_0 \beta_0$ значение $c_0 = 0$ даёт единственное решение уравнения (1.27). Для этой константы $W(c_0) = W(0) = \phi(0) = 1/\sqrt{2\pi}$, что доказывает совпадение (1.29) со второй частью (1.26).

Из равенства (1.27) следует, что $\beta_1 \Pi_1 - \beta_0 \Pi_0 = c_0 \Pi_0 \beta_0 / W(c_0)$. Это доказывает совпадение (1.29) с первой частью (1.26). \square

Замечание 7. В связи с формулой (1.29) интересно отметить особую роль ограничений β_0, β_1 на величину НОВ. В d-апостериорном подходе для регулярных вероятностных моделей необходимый объём выборки обратно пропорционален квадрату ограничений на d-риски. В то же время, в задаче различения простых гипотез с ограничениями на классические вероятности ошибок I-го и II-го рода НОВ слабо зависит от этих ограничений (скорость возрастания имеет порядок $O(\ln \beta)$) и основную роль играет расстояние между гипотезами — НОВ обратно пропорционален квадрату ширины области безразличия (см., например, [58]).

Проанализируем возможности применения схемы с жёсткими ограничениями в рамках рассмотренных выше моделей.

Модель N-N в схеме жёстких ограничений. В соответствии с (1.29) асимптотика НОВ при $\beta_0, \beta_1 \rightarrow 0$ (обозначается n_β) находится по формулам (1.27)–(1.29), в которых

$$\Pi_0 = 1 - \Pi_1 = \Phi\left(-\frac{\mu}{\tau}\right), \quad g(\theta_0) = \frac{1}{\tau}\phi\left(\frac{\mu}{\tau}\right), \quad I_0 = \sigma^{-2}.$$

Гипотеза H_0 принимается, когда выполняется неравенство (1.14), где константа c_0 находится из соотношения (1.27). Таким образом, $n^* \sim n_\beta$, где

$$n_\beta = \sigma^2 \left(\frac{W(c_0)\phi(\mu/\tau)}{\tau\Phi(-\mu/\tau)} \right)^2 \frac{1}{\beta_0^2}. \quad (1.31)$$

Асимптотическая формула (1.31) может быть уточнена. Способ такого уточнения основан на разложении в ряд Тейлора (при $\beta \rightarrow 0$, $n \rightarrow \infty$) в окрестности предельных значений аргументов всех частей формул (1.16), (1.17), определяющих НОВ. Для формулировки результата введём следующие обозначения:

$$\Delta_0 = \frac{\beta_1}{\beta_0} - \frac{\beta_0 + \beta_1}{\beta_0}\Pi_0, \quad Z = \Pi_0 + \Delta_0\Phi(c_0).$$

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

$$n^* = \left\lceil \sigma^2 \left(\frac{W(c_0)g_0}{\Pi_0} \frac{1}{\beta_0} + V(\beta_0) \right)^2 \right\rceil, \quad (1.32)$$

где $[a]$ — наименьшее целое число, не меньше a , и

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

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

Доказательство. Как уже отмечалось, достаточно ограничиться случаем $\sigma^2 = \tau^2 = 1$. Воспользовавшись равенством (1.5), представим безусловную вероятность принятия решения d_0 в виде

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

Снова в формулах (1.12)–(1.13) параметр n формально будем считать непрерывным — НОВ равно целой части (с округлением в большую сторону) того n , при котором в этих формулах достигается знак равенства. Обозначим $z = \sqrt{n}$. Из предыдущего представления в этом случае вместо (1.12)–(1.13) получаем систему уравнений

$$\Phi\left(\frac{t - \mu}{\sqrt{1 + 1/z^2}}\right) = \frac{\Phi(-\mu) - \rho\beta}{1 - (\rho + 1)\beta}, \quad (1.33)$$

$$\int_{-\infty}^{zt} \Phi(u) \phi\left(\frac{u}{z} - t + \mu\right) du = z\beta \frac{\Phi(-\mu) - \rho\beta}{1 - (\rho + 1)\beta}, \quad (1.34)$$

где $\beta = \beta_0$, $\beta_1 = \rho\beta$ и в формуле (1.12) в интеграле произведена замена переменной $\theta \rightarrow -u/z + t$, $u \in (-\infty, zt]$. Из (1.33) следует, что

$$t = \mu + \sqrt{1 + \frac{1}{z^2}} \Phi^{-1}(\varkappa),$$

где

$$\varkappa = \frac{\Phi(-\mu) - \rho\beta}{1 - (\rho + 1)\beta} = \Phi(-\mu) - \beta \Delta,$$

$$\Delta = \Delta(\beta) = \frac{\rho - \Pi_0(1 + \rho)}{1 - (1 + \rho)\beta} = \Delta_0 + \beta\zeta,$$

$$\Delta_0 = \rho - \Phi(-\mu)(1 + \rho), \quad \zeta = \Delta_0 \frac{1 + \rho}{1 - (1 + \rho)\beta}.$$

Легко видеть, что функция $\Phi^{-1}(x)$ в окрестности точки $x = x_0$ разлагается в ряд Тейлора (с остаточным членом в форме Лагранжа)

$$\Phi^{-1}(x) = \Phi^{-1}(x_0) + \frac{1}{\phi(\Phi^{-1}(x_0))} (x - x_0) + \frac{\Phi^{-1}(\tilde{x})}{2\phi^2(\Phi^{-1}(\tilde{x}))} (x - x_0)^2,$$

с $\tilde{x} \rightarrow x_0$ при $x \rightarrow x_0$. Воспользуемся этим разложением для $x_0 = \Phi(-\mu)$, $x = \varkappa$, а также разложением в ряд Тейлора функции $\sqrt{1+x}$ в точке $x = 0$. Получим (с $1 \geq \delta_z \rightarrow 1$, $\delta_\varkappa \rightarrow \mu$ при $\beta \rightarrow 0$, $z \rightarrow \infty$)

$$\begin{aligned} t &= \mu + \left(1 + \frac{1}{2z^2}\delta_z\right) \left(-\mu - \beta \frac{\Delta_0}{\phi(-\mu)} - \beta^2 \frac{\zeta}{\phi(-\mu)} - \beta^2 \frac{\Delta^2 \delta_\varkappa}{2\phi^2(-\delta_\varkappa)}\right) = \\ &= -\beta \frac{\Delta_0}{\phi(\mu)} - \frac{\mu}{2z^2}\delta_z - \beta^2 \left(\frac{\zeta}{\phi(\mu)} + \frac{\Delta^2 \delta_\varkappa}{2\phi^2(\mu)}\right) + O(\beta/z^2), \end{aligned} \quad (1.35)$$

где $\phi(-\mu) = \phi(\mu)$ в силу чётности стандартной нормальной функции плотности.

Из формулы (1.31) следует, что $z = A_1(\beta)/\beta$, где $\lim_{\beta \rightarrow 0} A_1(\beta) = A_1^*$ с соответствующей константой A_1^* (см. (1.31)). Поэтому при $\beta \rightarrow 0$ произведение $zt \rightarrow -A_1^* \Delta_0 / \phi(\mu) = c_0$. Подставляя это соотношение в (1.34) и переходя к пределу по $\beta \rightarrow 0$, получаем, что константа A_1^* должна удовлетворять уравнению

$$\phi(\mu) \int_{-\infty}^{c_0} \Phi(u) du = A_1^* \Phi(-\mu) = -c_0 \frac{\phi(\mu)}{\Delta_0} \Phi(-\mu)$$

(при $\Delta_0 = 0$ отношение $c_0/\Delta_0 = -A_1^*/\phi(\mu)$). В терминах функции (1.30)

$$W(c_0) = -c_0 \frac{\Phi(-\mu)}{\Delta_0}.$$

Заметим, что переход к пределу под знаком интеграла в (1.34) разрешён, т.к. при достаточно большом z верхний предел интегрирования $zt < C = c_0 + 1 < \infty$ и подынтегральная функция не превосходит интегрируемой на интервале $(-\infty, C]$ функции $\Phi(u)$. Кроме того, в силу (1.28) решение этого уравнения относительно A_1^* приводит к той же асимптотической формуле, что и (1.31) (с $A_1^* = W(c_0)\phi(\mu)/\Phi(-\mu)$).

Пусть теперь $z\beta = A_1^* + A_2(\beta)\beta$. Найдём $A_2^* = \lim_{\beta} A_2(\beta)$. Во-первых, из (1.35) следует, что при $\beta \rightarrow 0$

$$\begin{aligned} \frac{1}{\beta}(zt - c_0) &= -A_2 \frac{\Delta_0}{\phi(\mu)} - \frac{\mu}{2z\beta}\delta_z - z\beta \left(\frac{\zeta}{\phi(\mu)} + \frac{\Delta^2 \delta_\varkappa}{2\phi^2(\mu)}\right) + O(1/z) \rightarrow \\ &\rightarrow -A_2^* \frac{\Delta_0}{\phi(\mu)} - \frac{\mu}{2A_1^*} - \frac{A_1^* \Delta_0}{\phi(\mu)} \left(1 + \rho + \frac{\mu \Delta_0}{2\phi(\mu)}\right) = \Upsilon. \end{aligned}$$

Далее, представим обе части уравнения (1.34) в виде рядов по степеням β или $1/z$. В левой части уравнения воспользуемся представлением интеграла

(для любой интегрируемой функции h)

$$\int_{-\infty}^x h(u) du = \int_{-\infty}^{x_0} h(u) du + h(\tilde{x})(x - x_0),$$

где $\tilde{x} \rightarrow x_0$ при $x \rightarrow x_0$. Кроме того, для функции плотности ϕ запишем разложение

$$\phi\left(\frac{u}{z} - t + \mu\right) = \phi(\mu) + \left(\frac{u}{z} - t\right)\delta_u,$$

где $\delta_u \rightarrow \phi'(\mu) = -\mu\phi(\mu)$, причём $|\delta_u| < 1$. Таким образом, имеем равенство

$$\begin{aligned} & \int_{-\infty}^{zt} \Phi(u)\phi\left(\frac{u}{z} - t + \mu\right) du = \\ & = \phi(\mu) \int_{-\infty}^{c_0} \Phi(u) du + \frac{1}{z} \int_{-\infty}^{c_0} u\Phi(u)\delta_u du - t \int_{-\infty}^{c_0} \Phi(u)\delta_u du + \\ & \quad + (zt - c_0) \Phi(\delta) \phi\left(\frac{\delta}{z} - t + \mu\right), \end{aligned}$$

где $\delta \rightarrow c_0$. Опять же переход к пределу под знаком интеграла разрешён, т.к. $|\delta_u| < 1$ и область интегрирования ограничена сверху.

Первое слагаемое в последнем равенстве равно $A_1^*\Phi(-\mu)$. Напомним также, что $\delta_u \rightarrow -\mu\phi(\mu)$, $z\beta \rightarrow A_1^*$, $t/\beta = zt/(z\beta) \rightarrow c_0/A_1^* = -\Delta_0/\phi(\mu)$ и, кроме того, $(zt - c_0)/\beta \rightarrow \Upsilon$. Следовательно,

$$\begin{aligned} & \frac{1}{\beta} \left(\int_{-\infty}^{zt} \Phi(u)\phi\left(\frac{u}{z} - t + \mu\right) du - A_1^*\Phi(-\mu) \right) \rightarrow \\ & \rightarrow -\frac{\mu\phi(\mu)}{A_1^*} \int_{-\infty}^{c_0} u\Phi(u) du - \mu\Delta_0 \int_{-\infty}^{c_0} \Phi(u) du + \Upsilon \Phi(c_0)\phi(\mu) = \\ & = -\frac{\mu\phi(\mu)}{2A_1^*} ((c_0)^2\Phi(c_0) + c_0\phi(c_0) - \Phi(c_0)) - \frac{A_1^*\mu\Phi(-\mu)\Delta_0}{\phi(\mu)} + \Upsilon\Phi(c_0)\phi(\mu). \end{aligned}$$

Правая часть уравнения (1.34) равна

$$\begin{aligned} z\beta \frac{\Phi(-\mu) - \rho\beta}{1 - (\rho + 1)\beta} & = (A_1^* + A_2\beta) (\Phi(-\mu) - \beta\Delta_0 - \beta^2\zeta) = \\ & = A_1^*\Phi(-\mu) + (A_2\Phi(-\mu) - A_1^*\Delta_0)\beta + o(\beta). \end{aligned}$$

Следовательно,

$$\lim_{\beta \rightarrow 0} \frac{1}{\beta} \left(z\beta \frac{\Phi(-\mu) - \rho\beta}{1 - (\rho + 1)\beta} - A_1^*\Phi(-\mu) \right) = A_2^*\Phi(-\mu) - A_1^*\Delta_0.$$

Таблица 1.1: Сравнение асимптотических формул НОВ при малых β_0, β_1 ; N-N модель

	$\beta_1 = \beta_0 = 0.01$			$\beta_1 = \beta_0 = 0.001$		$\beta_0 = 0.001, \beta_1 = 0.005$				
μ	2	1	0	2	1	2	1	0	-1	-2
n^*	51	540	1 013	6 441	54 484	417	5 617	16 059	13 958	2 758
$n_\beta - n^*$	15	6	1	123	45	61	66	37	0	27
$\tilde{n}_\beta - n^*$	4	1	1	3	0	5	1	1	0	2

Отсюда и из предыдущего получаем уравнение для отыскания A_2^* :

$$\begin{aligned}
 & -\frac{\mu\phi(\mu)}{2A_1^*}((c_0)^2\Phi(c_0) + c_0\phi(c_0) - \Phi(c_0)) + c_0\mu\Phi(-\mu) + A_1^*\Delta_0 = \\
 & = A_2^*\Phi(-\mu) - \Upsilon\Phi(c_0)\phi(\mu) = \\
 & = A_2^*(\Phi(-\mu) + \Delta_0\Phi(c_0)) + \left(\frac{\mu}{2A_1^*} - c_0\left(1 + \rho + \frac{\mu\Delta_0}{2\phi(\mu)}\right)\right)\Phi(c_0)\phi(\mu).
 \end{aligned}$$

Требуемый результат получается простыми алгебраическими преобразованиями и подстановкой исключённых параметров σ, τ . \square

Замечание 8. В ситуации, соответствующей условиям теоремы 1.1, т.е. при $\beta_0 = \beta_1$ и априорном среднем $\mu = 0$, в предыдущей теореме коэффициент $Q = 0$. Поэтому формула (1.32) эквивалентна асимптотической формуле (1.31). Результат применения этих формул, хотя всего на одно наблюдение, но отличается от точного выражения (1.15) для НОВ в этом частном случае. Это связано, конечно, с различием в значениях $\text{tg}(\pi\beta_0)$ и $\pi\beta_0$. Чтобы учесть это различие, необходимо рассматривать представление n^* вида $\sigma^2(A_1/\beta + A_2 + A_3\beta)^2$. Для указанного здесь частного случая параметров в этом представлении $A_2 = 0$ и $A_3 = -\pi/3$.

В таблице 1.1 приведены ошибки приближений с помощью асимптотической формулы (1.31) (строка $n_\beta - n^*$) и с помощью асимптотической формулы (1.32) (строка $\tilde{n}_\beta - n^*$) при некоторых параметрах модели (различных значениях априорного среднего μ и различных ограничениях на d-риски; $\sigma^2 = \tau^2 = 1$). Как видно из таблицы, ошибка приближения n_β , может быть большой не только в абсолютном, но и в относительном выражении. Исправленная асимптотика значительно улучшает ситуацию.

Модель G-E в схеме с жёсткими ограничениями. Рассмотрим теперь случай с показательным распределением наблюдений и гамма-распределением параметра интенсивности (1.18). Как отмечено выше, граничную точку между гипотезами можно выбрать $\theta_0 = 1$. Асимптотика n_β НОВ при $\beta_0, \beta_1 \rightarrow 0$ находится по формулам (1.28), (1.29), где информация по Фишеру $I_0 = 1$,

$$\begin{aligned}\Pi_0 = 1 - \Pi_1 &= \mathbb{G}(a; \lambda, 1) = \frac{1}{\Gamma(\lambda)} \int_0^a \theta^{\lambda-1} e^{-\theta} d\theta, \\ g(\theta_0) &= g(1; \lambda, a) = \frac{a^\lambda}{\Gamma(\lambda)} e^{-a}.\end{aligned}$$

Таким образом,

$$n^* \sim n_\beta = \left[\left(\frac{W(c_0) g(1; \lambda, a)}{\mathbb{G}(a; \lambda, 1)} \frac{1}{\beta_0} \right)^2 \right]. \quad (1.36)$$

Гипотеза H_0 принимается, когда выборочное среднее

$$\begin{aligned}\bar{x} &\geq \frac{1}{n_\beta} [\mathbb{G}^{-1}(\Phi(-c_0); n_\beta + \lambda, 1) - a] \approx \\ &\approx 1 - \frac{\sqrt{n_\beta + \lambda}}{n_\beta} c_0 + \frac{\lambda - a + (c_0^2 - 1)/3}{n_\beta}.\end{aligned}$$

Константа c_0 находится из соотношения (1.28). Более короткий путь нахождения критической константы состоит в использовании равенства (1.20). Нулевую гипотезу следует принимать, если выборочное среднее \bar{x} удовлетворяет неравенству

$$\frac{\bar{x}}{\bar{x} + a/n_\beta} \geq \mathbb{B}^{-1}(Q; n_\beta, \lambda), \quad Q = \frac{1 - \Pi_0 - \beta_0}{1 - \beta_0 - \beta_1}.$$

Для того чтобы получить уточнение асимптотической формулы (1.36), положим $\beta = \beta_0$, $\rho = \beta_1/\beta_0$, и запишем уравнение (1.20) согласно хорошо известной формулы $\mathbb{B}(q; n, \lambda) = 1 - \mathbb{B}(1-q; \lambda, n)$ для функции распределения бета-закона в виде

$$\mathbb{B}\left(\frac{1}{t+1}; \lambda, n\right) = 1 - \frac{1 - \Pi_0 - \beta_0}{1 - \beta_0 - \beta_1} = \frac{\Pi_0 - \rho\beta}{1 - (1 + \rho)\beta} = \Pi_0 + \beta\Delta,$$

где

$$\Delta = \frac{\Pi_0 - \rho\Pi_1}{1 - (1 + \rho)\beta} = \varkappa + \varkappa(1 + \rho)\beta + O(\beta^2), \quad \beta \rightarrow 0, \quad (1.37)$$

с $\varkappa = \Pi_0 - \rho\Pi_1$.

Воспользуемся асимптотическим представлением для квантилей бета-распределения с $n \rightarrow \infty$, описанным в [3, стр. 29]. В соответствии с этим представлением имеем

$$\frac{1}{t+1} = \frac{\widehat{\chi}_\lambda^2}{\widehat{\chi}_\lambda^2/2 + 2n + \lambda - 1} + C_n, \quad (1.38)$$

где $\widehat{\chi}_\lambda^2$ — квантиль хи-квадрат распределения с 2λ степенями свободы порядка $\Pi_0 + \beta\Delta$; остаточный член $C_n = O(1/n^3)$. Учитывая связь хи-квадрат распределения с гамма-распределением, можно записать указанную квантиль через обратную функцию гамма-распределения в виде $\widehat{\chi}_\lambda^2 = 2\mathbb{G}^{-1}(\Pi_0 + \beta\Delta; \lambda, 1)$. Так как $\Pi_0 = \mathbb{G}(a; \lambda, 1)$, то

$$\lim_{\beta \rightarrow 0} \widehat{\chi}_\lambda^2 = 2\mathbb{G}^{-1}(\mathbb{G}(a; \lambda, 1); \lambda, 1) = 2a.$$

Следовательно, если $\beta \rightarrow 0$ и $n \rightarrow \infty$, то для константы t , дающей решение уравнения (1.20) (или (1.38)), имеем $t/n \rightarrow 1/a$.

Представим функцию $\Upsilon(\beta) = \mathbb{G}^{-1}(\Pi_0 + \beta\Delta; \lambda, 1)$, $\beta \in [0, 1]$, в виде отрезка ряда Маклорена. Заметим сначала, что $\Upsilon'(\beta) = \Delta/g(\Upsilon(\beta); \lambda, 1)$, $\Upsilon(0) = a$ и $g(a; \lambda, 1) = g(1; \lambda, a)/a$. Следовательно,

$$\left. \frac{d\Upsilon(\beta)}{d\beta} \right|_{\beta=0} = \frac{a\Delta}{g(1; \lambda, a)}.$$

Так как $g'(x; \lambda, 1) = dg(x; \lambda, 1)/dx = -(x+1-\lambda)g(x; \lambda, 1)/x$, то

$$\left. \frac{d^2\Upsilon(\beta)}{d\beta^2} \right|_{\beta=0} = -\frac{\Delta}{g^2(\Upsilon(\beta); \lambda, 1)} \left. \frac{g'(\Upsilon(\beta); \lambda, 1) \Delta}{g(\Upsilon(\beta); \lambda, 1)} \right|_{\beta=0} = \frac{a(a+1-\lambda)\Delta^2}{g^2(1; \lambda, a)}.$$

Таким образом, в соответствии с (1.37) справедливо представление

$$\begin{aligned} \widehat{\chi}_\lambda^2 &= 2a + \frac{2a\Delta}{g(1; \lambda, a)}\beta + \frac{a(a+1-\lambda)\Delta^2}{g^2(1; \lambda, a)}\beta^2 + O(\beta^3) = \\ &= 2a(1 + A\beta + B\beta^2) + O(\beta^3), \end{aligned}$$

где

$$\begin{aligned} A &= \frac{\varkappa}{g(1; \lambda, a)}, \\ B &= \frac{(a+1-\lambda)\Delta^2 + 2\varkappa(1+\rho)g(1; \lambda, a)}{2g^2(1; \lambda, a)} \xrightarrow{\beta \rightarrow 0} \frac{1}{2}(a+1-\lambda)A^2 + (1+\rho)A. \end{aligned}$$

Как и в модели N–N рассмотрим связь между НОВ и $\beta = \beta_0$ вида $\sqrt{n}\beta = h + h_1\beta$, т.е.

$$\beta = \frac{h}{\sqrt{n} - h_1} = \frac{h}{\sqrt{n}} + \frac{hh_1}{n} + o(1/n), \quad n \rightarrow \infty.$$

В этом случае при $\beta \rightarrow 0$ ($n \rightarrow \infty$) имеем

$$\hat{\chi}_\lambda^2 = 2a \left(1 + Ah \frac{1}{\sqrt{n}} + (Ahh_1 + Bh^2) \frac{1}{n} \right) + o(1/n).$$

Подставляя полученное выражение в уравнение (1.38) и разрешая это уравнение относительно $s = at/n$, получаем

$$s = 1 - Ah \frac{1}{\sqrt{n}} - \frac{1}{2}(1 + a - \lambda - 2A^2h^2 + 2Bh^2 + 2Ahh_1) \frac{1}{n} + o(1/n). \quad (1.39)$$

Положим в уравнении (1.19) $t = ns/a$ и осуществим замену переменной интегрирования $\theta \rightarrow a(1 + u/\sqrt{n})/s$, $u \in (-\sqrt{n}, (s-1)\sqrt{n})$. Кроме того, для сокращения записи обозначим $g(\theta) = g(\theta; \lambda, 1)$. В результате получим уравнение (относительно переменной s)

$$\frac{a}{s\sqrt{n}} \int_{-\sqrt{n}}^{(s-1)\sqrt{n}} \mathbb{G}(n + u\sqrt{n}; n, 1) g\left(\frac{a}{s} + \frac{au}{s\sqrt{n}}\right) du = \rho\beta(\Pi_1 - \beta\Delta).$$

Заменим здесь функцию гамма-распределения на ряд Эджворта (1.22) с остаточным членом $C_n(u) = \phi(u)P(u)/\sqrt{n} + R_n(u)/\sqrt{n}$, где $P(u)$ — полином u и $\sup_u |R_n(u)| \rightarrow 0$ (см. [24, Гл. XVI, § 4]). Заметим, что

$$\left| \int_{-\sqrt{n}}^{(s-1)\sqrt{n}} R_n(u) g\left(\frac{a}{s} + \frac{au}{s\sqrt{n}}\right) du \right| \frac{1}{n} \leq \sup_u |R_n(u)| \int_0^1 g(\theta) d\theta \frac{s}{a\sqrt{n}} = o\left(\frac{1}{\sqrt{n}}\right).$$

Кроме того,

$$\begin{aligned} \left| \int_{-\sqrt{n}}^{(s-1)\sqrt{n}} \phi(u)P(u) g\left(\frac{a}{s} + \frac{au}{s\sqrt{n}}\right) du \right| \frac{1}{n} &\leq \sup_{t>a/2} g(t) \int_{(s/2-1)\sqrt{n}}^{(s-1)\sqrt{n}} \phi(u)|P(u)| du \frac{1}{n} + \\ &+ \sup_{u \leq (s/2-1)\sqrt{n}} \phi(u)|P(u)| \int_{-\sqrt{n}}^{(s/2-1)\sqrt{n}} g\left(\frac{a}{s} + \frac{au}{s\sqrt{n}}\right) du \frac{1}{n} \leq \\ &\leq \sup_{t>a/2} g(t) \int_0^\infty \phi(u)|P(u)| du \frac{1}{n} + \sup_{u \leq (s/2-1)\sqrt{n}} \phi(u)|P(u)| \int_0^a g(\theta) d\theta \frac{1}{a\sqrt{n}} = \\ &= o\left(\frac{1}{\sqrt{n}}\right), \end{aligned}$$

где последнее равенство следует из того, что для плотности гамма-распределения $\sup_{t>a/2} g(t) < \infty$ и для достаточно больших n разность $s/2 - 1 < -1/4$, поэтому $\sup_{u \leq (s/2-1)\sqrt{n}} \phi(u) |P(u)| \xrightarrow[n]{n} 0$.

Таким образом, уравнение (1.19) можно переписать в виде

$$a \int_{-\sqrt{n}}^{(s-1)\sqrt{n}} \left(\Phi(u) + \phi(u) \frac{1-u^2}{3} \frac{1}{\sqrt{n}} \right) g\left(\frac{a}{s} + \frac{au}{s\sqrt{n}}\right) du + o\left(\frac{1}{\sqrt{n}}\right) = \quad (1.40)$$

$$= s\rho\sqrt{n}\beta(\Pi_1 - \beta\Delta).$$

Перейдём формально к пределу по $n \rightarrow \infty$, $\beta \rightarrow 0$ (возможность осуществления такого перехода будет видна из дальнейших выкладок; напомним ещё, что $\sqrt{n}\beta \rightarrow h \in (0, \infty)$). Из (1.39) следует, что $(s-1)\sqrt{n} \rightarrow -Ah$, кроме того, $ag(a) = g(1; \lambda, a)$, поэтому после перехода к пределу соотношение (1.40) принимает вид

$$\int_{-\infty}^{-Ah} \Phi(u) du = \frac{\Pi_1 \rho h}{g(1; \lambda, a)}. \quad (1.41)$$

Полагая $c_0 = Ah = h(\Pi_0 - \rho\Pi_1)/g(1; \lambda, a)$, получаем уравнение $W(-c_0) = c_0\rho\Pi_1/(\Pi_0 - \rho\Pi_1)$, которое в силу (1.27) эквивалентно уравнению (1.28) и значение $h = \lim_{\beta \rightarrow 0} \sqrt{n}\beta$ совпадает с (1.36).

Уточним равенство (1.40) с учётом полученного представления для s . Заметим сначала, что для достаточно больших n разность $s/2 - 1 < 1/4$, поэтому при $n \rightarrow \infty$

$$\int_{-\sqrt{n}}^{(s/2-1)\sqrt{n}} \left(\Phi(u) + \phi(u) \frac{1-u^2}{3} \frac{1}{\sqrt{n}} \right) g\left(\frac{a}{s} + \frac{au}{s\sqrt{n}}\right) du \leq$$

$$\leq \sup_{u < -\sqrt{n}/4} \left| \Phi(u) + \phi(u) \frac{1-u^2}{3} \frac{1}{\sqrt{n}} \right| \frac{s\sqrt{n}}{a} \int_0^{a/2} g(\theta) d\theta = o\left(\frac{1}{n}\right),$$

т.к. функции Φ и ϕ имеют экспоненциальную скорость убывания на $-\infty$.

Далее, представим функцию $g(\theta) = \theta^{\lambda-1}e^{-\theta}/\Gamma(\lambda)$, $\theta \geq a/2$, в виде отрезка ряда Тейлора относительно точки $\theta = a$:

$$g(\theta) = g(a) - \frac{g(a)}{a}(1+a-\lambda)(\theta-a) + r(\theta)(\theta-a)^2,$$

где для некоторого $\delta > 0$ функция $|r(\theta)| \leq \delta e^{-\theta/2}$ при $\theta \geq a/2$. Из (1.39) имеем $1/s = 1 + Ah/\sqrt{n} + O(1/n)$, $n \rightarrow \infty$. Следовательно,

$$\frac{a}{s} + \frac{au}{s\sqrt{n}} - a = a(Ah + u) \frac{1}{\sqrt{n}} + uO\left(\frac{1}{n}\right) + O\left(\frac{1}{n}\right)$$

и

$$g\left(\frac{a}{s} + \frac{au}{s\sqrt{n}}\right) = g(a) - g(a)(1+a-\lambda)(Ah+u)\frac{1}{\sqrt{n}} + r_n(u)\frac{1}{n},$$

где $|r_n(u)| \leq \delta V(u/\sqrt{n})e^{-u/2\sqrt{n}}$ и V есть полином степени, не выше второй.

Таким образом,

$$\begin{aligned} & a \int_{(s/2-1)\sqrt{n}}^{(s-1)\sqrt{n}} \Phi(u) g\left(\frac{a}{s} + \frac{au}{s\sqrt{n}}\right) du = \\ & = a \int_{(s/2-1)\sqrt{n}}^{(s-1)\sqrt{n}} \Phi(u) \left(g(a) - g(a)(1+a-\lambda)(Ah+u)\frac{1}{\sqrt{n}} + r_n(u)\frac{1}{n} \right) du. \end{aligned}$$

Первое слагаемое здесь (в силу (1.41) и теоремы о среднем)

$$\begin{aligned} a \int_{(s/2-1)\sqrt{n}}^{(s-1)\sqrt{n}} \Phi(u) g(a) du &= \int_{-\infty}^{-Ah} + \int_{-Ah}^{(s-1)\sqrt{n}} - \int_{-\infty}^{(s/2-1)\sqrt{n}} \Phi(u) ag(a) du = \\ &= \Pi_1 \rho h - \frac{1}{2} \Phi(\tilde{s}_n) \frac{ag(a)A_1}{\sqrt{n}} + O\left(\frac{1}{n}\right), \end{aligned}$$

где «средняя точка» $\tilde{s}_n \rightarrow -Ah$ и в соответствии с (1.39) $A_1 = (1+a-\lambda - 2A^2h^2 + 2Bh^2 + 2Ahh_1)$. Аналогично для второго слагаемого:

$$\begin{aligned} & -ag(a)(a+1-\lambda) \int_{(s/2-1)\sqrt{n}}^{(s-1)\sqrt{n}} \Phi(u)(Ah+u) du \frac{1}{\sqrt{n}} = \\ & = -\left(\Pi_1 \rho h^2 A + ag(a) \int_{-\infty}^{-Ah} u \Phi(u) du \right) \frac{(1+a-\lambda)}{\sqrt{n}} + O\left(\frac{1}{n}\right). \end{aligned}$$

Заметим, что $\int_{-\infty}^{-Ah} u \Phi(u) du = \frac{1}{2}(\Phi(-Ah)(A^2h^2 - 1) - \phi(Ah)Ah)$.

Наконец, третье слагаемое

$$a \int_{(s/2-1)\sqrt{n}}^{(s-1)\sqrt{n}} \Phi(u) r_n(u) du \frac{1}{n} = O\left(\frac{1}{n}\right),$$

т.к. интегральное выражение здесь ограничено в силу свойств функции $r_n(u)$.

Аналогично, имеем

$$\begin{aligned} & \frac{a}{\sqrt{n}} \int_{(s/2-1)\sqrt{n}}^{(s-1)\sqrt{n}} \phi(u) \frac{1-u^2}{3} g\left(\frac{a}{s} + \frac{au}{s\sqrt{n}}\right) du = \int_{(s/2-1)\sqrt{n}}^{(s-1)\sqrt{n}} \phi(u) \left(ag(a) \frac{1}{\sqrt{n}} - \right. \\ & \quad \left. - ag(a)(a+1-\lambda)(Ah+u)\frac{1}{n} + r_n(u)\frac{1}{n\sqrt{n}} \right) \frac{1-u^2}{3} du = \\ & = ag(a) \int_{-\infty}^{-Ah} \phi(u) \frac{1-u^2}{3} du \frac{1}{\sqrt{n}} + O\left(\frac{1}{n}\right) = -\frac{1}{3} ag(a) Ah \phi(Ah) + O\left(\frac{1}{n}\right). \end{aligned}$$

Таблица 1.2: Точность асимптотики n_β для G–E модели

		$\lambda = 1,$ $a = 0.5$	$\lambda = 1,$ $a = 1$	$\lambda = 5,$ $a = 5$	$\lambda = 3,$ $a = 4$	$\lambda = 12,$ $a = 9$
$\beta_0 = 0.10$	n^*	6	10	48	26	49
$\beta_1 = 0.10$	$n_\beta - n^*$	1	0	2	1	16
	$\tilde{n}_\beta - n^*$	0	0	2	2	9
$\beta_0 = 0.05$	n^*	24	37	196	107	236
$\beta_1 = 0.05$	$n_\beta - n^*$	1	0	2	0	22
	$\tilde{n}_\beta - n^*$	0	0	2	2	9
$\beta_0 = 0.01$	n^*	600	906	4947	2676	6361
$\beta_1 = 0.01$	$n_\beta - n^*$	4	-4	-2	-7	70
	$\tilde{n}_\beta - n^*$	0	0	1	1	10
$\beta_0 = 0.001$	n^*	3164	6966	33907	26692	24151
$\beta_1 = 0.01$	$n_\beta - n^*$	68	72	305	127	692
	$\tilde{n}_\beta - n^*$	1	1	5	2	17

Следовательно, левая часть (1.40) равна

$$\Pi_1 \rho h + \frac{1}{\sqrt{n}} A^* + o\left(\frac{1}{\sqrt{n}}\right),$$

где

$$\begin{aligned} \lim_{\beta \rightarrow 0} A^* = & \left[-Ah^2 \Pi_1 \rho - ag(a) \frac{1}{2} (\Phi(-Ah)(A^2 h^2 - 1) - \phi(Ah)Ah) \right] (a + 1 - \lambda) - \\ & - ag(a) \Phi(-Ah) - \frac{1}{2} ag(a) A_1 \Phi(-Ah). \end{aligned}$$

Правая часть (1.40) равна

$$\Pi_1 \rho h + (h\rho \Pi_1 h_1 - h^2 \Delta \rho - Ah^2 \Pi_1 \rho) \frac{1}{\sqrt{n}} + O\left(\frac{1}{n}\right),$$

где $\Delta \rightarrow ag(a)A = \Pi_0 - \rho \Pi_1$. Таким образом, нами доказана

Теорема 1.4. *Для гамма-показательной модели с априорной гамма-плотностью $g(\theta; \lambda, a)$ необходимый объём выборки при различении гипотез $H_0: \theta \leq 1$, $H_1: \theta > 1$ и ограничениях $\beta_1 = \rho \beta_0 \rightarrow 0$ имеет асимптотическое представление*

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

где $g_0 = g(\theta_0; \lambda, a)$,

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

Таблица 1.3: Точность асимптотики n_β для В-В модели

		$p = 0.1,$ $q = 10,$ $\theta_0 = 0.01$	$p = 1,$ $q = 1,$ $\theta_0 = 0.5$	$p = 1,$ $q = 1,$ $\theta_0 = 0.3$	$p = 5,$ $q = 5,$ $\theta_0 = 0.5$	$p = 5,$ $q = 5,$ $\theta_0 = 0.4$	$p = 10,$ $q = 10,$ $\theta_0 = 0.5$
$\beta_0 = 0.10$	n^*	31	15	15	91	71	185
$\beta_1 = 0.10$	$n_\beta - n^*$	20	1	0	6	7	13
$\beta_0 = 0.05$	n^*	153	63	60	379	303	779
$\beta_1 = 0.05$	$n_\beta - n^*$	50	1	0	7	9	12
$\beta_0 = 0.01$	n^*	4877	1591	1495	9633	7767	19751
$\beta_1 = 0.01$	$n_\beta - n^*$	196	1	-3	6	17	14

$$Q_1 = -\frac{\Delta_0}{6}(3a\theta_0 + 1 - 3\lambda)\phi(c_0) + \rho(\theta_0 g_0 - (a\theta_0 - \lambda)(1 - \Pi_0))c_0 - \\ - \theta_0 g_0(1 + \rho)\Phi(-c_0)c_0 + \Delta_0(a\theta_0 - \lambda)\Phi(-c_0)c_0,$$

$$Z = \Pi_0 + \Delta_0\Phi(c_0) = \Pi_0 + (\rho - (1 + \rho)\Pi_0)\Phi(c_0).$$

Из таблицы 1.2 видно, что предложенная асимптотическая формула для НОВ (в таблице ошибка формулы обозначена $\tilde{n}_\beta - n^*$) значительно улучшает точность аппроксимации.

Модель В-В в схеме с жёсткими ограничениями. Для модели бета-бернулли (1.23) асимптотика n_β НОВ при малых β_0, β_1 находится по формулам (1.28), (1.29), где информация по Фишеру $I_0 = 1/(\theta_0(1 - \theta_0))$ и

$$\Pi_0 = 1 - \Pi_1 = \mathbb{B}(\theta_0; p, q) = \frac{1}{\mathbb{B}(p, q)} \int_0^{\theta_0} x^{p-1}(1-x)^{q-1} dx, \\ g(\theta_0) = g(\theta_0; p, q) = \frac{1}{\mathbb{B}(p, q)} \theta_0^{p-1}(1-\theta_0)^{q-1}.$$

Гипотеза H_0 принимается, когда апостериорная вероятность нулевой гипотезы $\mathbb{B}(\theta_0; p + S_n, q + n - S_n) \geq \Phi(c_0)$, где константа c_0 находится из соотношения (1.28) и S_n есть сумма выборочных значений.

Из таблицы 1.3 видно, что асимптотическое приближение n^* в схеме с жёсткими ограничениями обладает почти идеальной точностью.

1.2.2 Асимптотика необходимого объёма выборки в схеме стягивающегося априори

Другой подход к выводу асимптотической формулы для НОВ применяет-

ся в том случае, когда априорное распределение имеет большую концентрацию на границе θ_0 различаемых гипотез (так называемая схема «стягивающегося априори», см. [8]). В этом случае функция плотности априорного распределения представляется в виде $g(\theta) = g((\theta - \theta_0)/\tau; \tau)/\tau$ с некоторым вспомогательным параметром $\tau > 0$ и асимптотика НОВ осуществляется для $\tau \rightarrow 0$. При этом предполагается, что $\lim_{\tau} g(u; \tau) = g(u; 0)$. При определённых условиях регулярности (см. [8], стр. 63, а также [58]), которые выполняются для байесовских моделей, рассматриваемых в диссертации, асимптотика НОВ задаётся следующим соотношением:

$$n^* \sim n_{\tau} = \frac{r^2}{I_0} \frac{1}{\tau^2}, \quad (1.42)$$

где константа r есть решение системы уравнений (относительно r, s)

$$\begin{aligned} \int_0^{\infty} \Phi(r(s-u)) g(u; 0) du &= \beta_0 \frac{\tilde{\Pi}_0 - \beta_1}{1 - \beta_0 - \beta_1}, \\ \int_{-\infty}^0 \Phi(r(u-s)) g(u; 0) du &= \beta_1 \frac{\tilde{\Pi}_1 - \beta_0}{1 - \beta_0 - \beta_1}, \end{aligned} \quad (1.43)$$

$\tilde{\Pi}_0 = 1 - \tilde{\Pi}_1 = \int_{-\infty}^0 g(u; 0) du$, Φ — функция распределения стандартного нормального $(0, 1)$ закона. Соответствующая критическая константа C для правила вида (1.6) имеет конечное предельное значение C_0 :

$$\lim_{\tau \rightarrow 0} C = C_0 = \frac{\int_{-\infty}^0 \phi(r(s-u)) g(u; 0) du}{\int_{-\infty}^{\infty} \phi(r(s-u)) g(u; 0) du}, \quad (1.44)$$

где ϕ — функция плотности стандартного нормального $(0, 1)$ закона.

Часто параметр τ может быть выбран так, что $g(u; 0) = \phi(u - \mu_0)$ с некоторым $\mu_0 \in \mathbb{R}^1$. В этом случае отношение интегралов в формуле для C_0 равно $\Phi(- (sr^2 + \mu_0)/\sqrt{1+r^2})$. То есть в этом случае константа

$$C_0 = \Phi(-c_0), \quad \text{где } c_0 = \frac{sr^2 + \mu_0}{\sqrt{r^2 + 1}}. \quad (1.45)$$

Система (1.43) также может быть записана в более простом виде (подробности см. в следующем разделе). Если же, кроме того, параметр $\mu_0 = 0$ и рассматриваются совпадающие ограничения на d-риски $\beta_0 = \beta_1 = \beta$, то правые части

уравнений (1.43) совпадают (равны $\beta/2$) и константа $s = 0$ (при $s = 0$ левые части (1.43) также будут совпадать). Интеграл в левой части (1.43) при $s = 0$ равен $\text{arctg}(1/r)/2\pi$. Следовательно, предельное значение критической константы $c_0 = 0$ ($C_0 = 1/2$) и приближённое значение НОВ

$$n_\tau = \frac{1}{I_0} \frac{1}{\tau^2 \text{tg}^2(\pi\beta)}. \quad (1.46)$$

Модель N-N в схеме стягивающегося априори. Рассмотрим задачу различения двух односторонних гипотез $H_0 : \theta \leq \theta_0$ и $H_1 : \theta > \theta_0$ о среднем значении $\theta \in \mathbb{R}^1$ нормального $\mathcal{N}(\theta, \sigma^2)$ распределения с известной дисперсией σ^2 , где θ есть реализация случайной величины ϑ с априорным нормальным (μ, τ^2) распределением (нормально-нормальная модель (1.10)). Покажем, что формулы (1.42)–(1.46), предназначенные для приближённого вычисления необходимого объёма выборки n^* в случае, когда $\mu \approx \theta_0$ и $\tau \rightarrow 0$, здесь дают точное значение n^* .

Лемма 1.8. *Для нормально-нормальной модели (1.10) формулы (1.42)–(1.46) дают способ вычисления точного значения НОВ n^* .*

Действительно, рассуждая аналогично доказательству теоремы 1.1, приходим к тому, что n^* можно найти как округлённое вверх значение «непрерывного» параметра n , удовлетворяющего системе (1.12)–(1.13), где во втором соотношении достигается знак равенства. Воспользовавшись тождеством (1.5), получим, что безусловная вероятность принятия решения d_0 удовлетворяет уравнению

$$\Phi\left(\frac{t - \mu}{\sqrt{\tau^2 + \sigma^2/n}}\right) = \frac{\Pi_0 - \mathcal{R}_1(n, t)}{1 - \mathcal{R}_0(n, t) - \mathcal{R}_1(n, t)} = \frac{\Pi_0 - \beta_1}{1 - \beta_0 - \beta_1}. \quad (1.47)$$

Подставляя это соотношение в (1.12)–(1.13), замечаем, что после замены переменной интегрирования $\theta/\tau \dashrightarrow u$, а также после переобозначения параметров $t/\tau \dashrightarrow s$, $\mu/\tau \dashrightarrow \mu_0$, $n\tau^2/\sigma^2 \dashrightarrow r^2$, система (1.43) с функцией плотности $g(u; 0) = \phi(u - \mu_0)$ полностью совпадает с системой уравнений (1.12)–(1.13). При этом, как и заявлено в (1.42) НОВ $n^* \approx r^2\sigma^2/\tau^2$.

Заменяя в соотношении (1.14) параметры их выражениями из (1.42)–(1.46), получаем, что гипотеза $H_0 : \theta \leq 0$ должна приниматься, если вычисленное по выборке объёма $n = \lceil r^2\sigma^2/\tau^2 \rceil$ выборочное среднее $\bar{x} \leq s\tau$. В

данном случае, однако, удобнее представить область принятия нулевой гипотезы, воспользовавшись равенством (1.47), т.е. гипотезу $H_0 : \theta \leq 0$ следует принимать в случае, когда выборочное среднее

$$\bar{x} \leq \mu + \sqrt{\tau^2 + \frac{\sigma^2}{n}} \Phi^{-1} \left(\frac{\Pi_0 - \beta_1}{1 - b_0 - b_1} \right).$$

Легко показать, что оба варианта представления области принятия нулевой гипотезы совпадают между собой.

Если $\mu_0 = 0$ и $\beta_0 = \beta_1 = \beta$, то асимптотическая формула (1.46) совпадает с точной формулой теоремы 1.1. Константа $s = 0$, поэтому критерий принимает нулевую гипотезу, если $\bar{x} \leq 0$.

Модель G-E в схеме стягивающегося априори. Рассмотрим теперь задачу различения гипотез $H_0 : \theta \leq \theta_0$ и $H_1 : \theta > \theta_0$ в гамма-показательной модели (1.18) с граничной точкой $\theta_0 = 1$.

Как известно, среднее значение и дисперсия априорного гамма-распределения $\mathcal{G}(\lambda, a)$ равны λ/a и λ/a^2 соответственно. Следовательно, концентрация этого распределения около точки $\theta_0 = 1$ возможна только тогда, когда $\lambda/a \rightarrow 1$, $\lambda/a^2 \rightarrow 0$. Воспользовавшись формулой Стирлинга $\Gamma(\lambda) = \delta \sqrt{2\pi} \lambda^\lambda \cdot \exp\{-\lambda\}/\sqrt{\lambda}$, $\delta \rightarrow 1$, представим плотность гамма-распределения в следующем виде:

$$g(\theta; \lambda, a) = \frac{\sqrt{\lambda}}{\sqrt{2\pi}} \left(\frac{a}{\lambda}\right)^\lambda \exp\{\lambda - a\theta + (\lambda - 1) \ln(\theta)\} \delta.$$

Положим здесь $a = \lambda - \omega$ и $(\theta - 1)/\tau = x$. Получим

$$g(\theta; \lambda, a) = \frac{1}{\tau} h(x; \tau) e^{x\omega\tau} \left(1 - \frac{\omega}{\lambda}\right)^\lambda e^\omega \delta,$$

где

$$h(x; \tau) = \frac{1}{\sqrt{2\pi}} \exp\{-\lambda\tau x + (\lambda - 1) \ln(1 + \tau x)\}.$$

При $\tau = 1/\sqrt{\lambda}$, $\lambda \rightarrow \infty$, предел выражения под знаком экспоненты равен $-x^2/2$. Действительно, по формуле Тейлора для функции $\ln(1 + \tau x)$ имеем

$$\begin{aligned} & -\lambda\tau x + (\lambda - 1) \ln(1 + \tau x) = \\ & = -\sqrt{\lambda}x + (\lambda - 1) \left(\frac{1}{\sqrt{\lambda}} x - \frac{1}{2} \frac{1}{\lambda} x^2 + O\left(\frac{1}{\sqrt{\lambda^3}}\right) \right) = -\frac{1}{2} x^2 + O\left(\frac{1}{\sqrt{\lambda}}\right). \end{aligned}$$

Если, кроме того, $\omega\tau = \omega/\sqrt{\lambda} \rightarrow \mu$ с некоторым $\mu \in (-\infty, \infty)$, то $x\omega\tau \rightarrow x\mu$ и в силу той же формулы Тейлора

$$\left(1 - \frac{\omega}{\lambda}\right)^\lambda e^\omega = \exp\left\{\omega - \frac{\omega^2}{2\lambda} + o(\omega^2/\lambda)\right\} \rightarrow e^{-\mu^2/2}.$$

Таким образом, имеет место следующая

Лемма 1.9. Пусть параметры гамма-распределения $\mathcal{G}(\lambda, a)$ удовлетворяют условиям $\lambda \rightarrow \infty$, $(\lambda - a)/\sqrt{\lambda} \rightarrow \mu \in \mathbb{R}^1$. Тогда плотность этого распределения допускает представление

$$g(\theta; \lambda, a) = \frac{1}{\tau} \tilde{g}\left(\frac{\theta - 1}{\tau}; \tau\right),$$

где $\tau = 1/\sqrt{\lambda}$ и при любом $x \in \mathbb{R}^1$

$$\lim_{\tau \rightarrow 0} \tilde{g}(x; \tau) = \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2} = \phi(x - \mu).$$

Замечание 9. Если не учитывать условий этой леммы и применять асимптотическую формулу с «предельной» плотностью $\tilde{g}(x; 0) = \phi(x)$ в ситуации, когда $\lambda \neq a$, то ошибка асимптотического приближения может оказаться сравнимой с НОВ.

Информация по Фишеру показательной модели $I(\theta_0) = 1/\theta_0^2 = 1$. Таким образом, чтобы воспользоваться асимптотикой n_τ для НОВ в схеме стягивающегося априори (т.е. при $\lambda/a \approx 1$, $\lambda/a^2 \ll 1$), необходимо в формулах (1.42)–(1.46) положить $g(u; 0) = \phi(u - \mu)$, $\mu = (\lambda - a)/\sqrt{\lambda}$ и $I_0 = 1$. Предельные значения априорных вероятностей гипотез $\tilde{\Pi}_0 = 1 - \tilde{\Pi}_1 = \Phi(-\mu)$. Необходимый объём выборки (при $\lambda \rightarrow \infty$, $(\lambda - a)/\sqrt{\lambda} \rightarrow \mu \in \mathbb{R}^1$)

$$n^* \sim n_\tau = r^2 \lambda.$$

Критерий принимает гипотезу $H_0 : \theta < 1$, когда апостериорная вероятность нулевой гипотезы $\mathbb{G}(a + n_\tau \bar{x}; n_\tau + \lambda, 1) > C$ (см. (1.44)). Таким образом, для представления области принятия в стандартном виде относительно суммы выборочных значений необходимо иметь формулу для вычисления обратной функции гамма-распределения.

Таблица 1.4: Точность асимптотики n_τ для G–E модели

		$\lambda = 1,$ $a = 0.5$	$\lambda = 1,$ $a = 1$	$\lambda = 5,$ $a = 5$	$\lambda = 3,$ $a = 4$	$\lambda = 12,$ $a = 9$
$\beta_0 = 0.10$	n^*	6	10	48	26	49
$\beta_1 = 0.10$	$n_\tau - n^*$	2	0	0	-3	14
$\beta_0 = 0.05$	n^*	24	37	196	107	236
$\beta_1 = 0.05$	$n_\tau - n^*$	10	3	4	-10	54
$\beta_0 = 0.01$	n^*	600	906	4947	2676	6361
$\beta_1 = 0.01$	$n_\tau - n^*$	271	107	116	-196	1265
$\beta_0 = 0.001$	n^*	3164	6966	33907	26692	24151
$\beta_1 = 0.01$	$n_\tau - n^*$	876	-624	-2198	-4013	4620

Воспользовавшись асимптотической формулой (1.21), получаем, что нулевая гипотеза должна приниматься, когда выборочная сумма

$$\sum_{i=1}^n x_i > n + t_C \sqrt{n + \lambda} + \lambda - a + \frac{1}{3}(t_C^2 - 1), \quad t_C = \Phi^{-1}(C). \quad (1.48)$$

В таблице 1.4 приведены некоторые результаты применения этого асимптотического приближения НОВ. Видно, что при умеренных различиях в параметрах λ и a и не слишком жёстких ограничениях асимптотика n_τ даёт хорошее приближение к n^* . В ситуациях, когда априорное среднее далеко от граничной точки $\theta_0 = 1$ (т.е. при $\lambda/a \neq 1$), а также при жёстких ограничениях на величину того или иного риска, абсолютная ошибка достигает больших значений.

Модель В-В в схеме стягивающегося априори. Пусть требуется проверить гипотезу $H_0 : \theta \leq \theta_0$ против альтернативы $H_1 : \theta > \theta_0$ в модели бета-Бернулли (1.23). Так как среднее значение бета-распределения равно $p/(p+q)$, а дисперсия $pq/((p+q)^2(1+p+q))$, то асимптотические формулы в схеме стягивающихся априори могут быть применены, если

$$\mu = \frac{p}{p+q} \rightarrow \theta_0, \quad \tau^2 = \frac{pq}{(p+q)^3} \rightarrow 0.$$

Ясно, что должно быть $p, q \rightarrow \infty$ одновременно.

Лемма 1.10. Пусть параметры априорного бета-закона $p, q \rightarrow \infty$ так, что

$$\mu = \frac{p}{p+q} \rightarrow \theta_0, \quad \frac{\mu - \theta_0}{\tau} \rightarrow \mu_0,$$

где $\tau^2 = pq/(p+q)^3$ и константа $\mu_0 \in \mathbb{R}^1$. Тогда априорная плотность может быть представлена в виде

$$g(\theta; p, q) = \frac{1}{\tau} \tilde{g}\left(\frac{\theta - \theta_0}{\tau}; \tau\right),$$

где $\tilde{g}(u; \tau) \rightarrow \phi(u - \mu_0)$ для любого $u \in \mathbb{R}^1$, ϕ — нормальная $(0, 1)$ функция плотности.

Доказательство. По формуле Стирлинга

$$B(p, q) = b_{p,q} \frac{\sqrt{2\pi}(p+q)^{p+q} \sqrt{p} e^p \sqrt{q} e^q}{\sqrt{p+q} e^{p+q} \sqrt{2\pi p^p} \sqrt{2\pi q^q}} = b_{p,q} \frac{1}{\sqrt{2\pi\tau}} \mu^{1-p} (1-\mu)^{1-q},$$

где мы воспользовались равенством $1 - \mu = q/(p+q)$; множитель $b_{p,q} \rightarrow 1$. Следовательно, априорная плотность

$$g(\theta; p, q) = b_{p,q} \frac{1}{\sqrt{2\pi\tau}} \left(\frac{\theta}{\mu}\right)^{p-1} \left(\frac{1-\theta}{1-\mu}\right)^{q-1}.$$

Положим в этой формуле $\theta = u\tau + \theta_0$. Получим

$$g(\theta; p, q) = \tilde{b}_{p,q} \frac{1}{\sqrt{2\pi\tau}} \exp\{p \ln(1 + U_p) + q \ln(1 - U_q)\},$$

где $\tilde{b}_{p,q} = b_{p,q} (\mu/(u\tau + \theta_0)) (1-\mu)/(1-u\tau - \theta_0) \rightarrow 1$ для любого $u \in \mathbb{R}^1$, и

$$U_p = \frac{\tau}{\mu} \left(u - \frac{\mu - \theta_0}{\tau}\right), \quad U_q = \frac{\tau}{1-\mu} \left(u - \frac{\mu - \theta_0}{\tau}\right).$$

Так как $\tau \rightarrow 0$, то с помощью разложения $\ln(1 \pm \varepsilon) = \pm \varepsilon - \frac{1}{2} \varepsilon^2 + o(\varepsilon^2)$ выражение под знаком экспоненты преобразуется к виду

$$pU_p - qU_q - \frac{1}{2}(pU_p^2 + qU_q^2) + o(\tau^2).$$

Так как $p/\mu = q/(1-\mu) = p+q$, то $pU_p - qU_q = 0$. Кроме того,

$$pU_p^2 + qU_q^2 = \left(u - \frac{\mu - \theta_0}{\tau}\right)^2 \left(\frac{p\tau^2}{\mu^2} + \frac{q\tau^2}{(1-\mu)^2}\right) = \left(u - \frac{\mu - \theta_0}{\tau}\right)^2,$$

что завершает доказательство. □

Следовательно, в условиях этой леммы априорная плотность может быть приближена плотностью $\phi((\theta - \theta_0)/\tau - \mu_0)/\tau$, т.е. плотностью нормального (μ, τ^2) закона (здесь учтено, что $\mu = p/(p+q) = \theta_0 + \tau\mu_0$). Как отмечено в

Таблица 1.5: Точность асимптотики n_τ для В-В модели

		$p = 0.1,$ $q = 10,$ $\theta_0 = 0.01$	$p = 1,$ $q = 1,$ $\theta_0 = 0.5$	$p = 1,$ $q = 1,$ $\theta_0 = 0.3$	$p = 5,$ $q = 5,$ $\theta_0 = 0.5$	$p = 5,$ $q = 5,$ $\theta_0 = 0.4$	$p = 10,$ $q = 10,$ $\theta_0 = 0.5$
$\beta_0 = 0.10$	n^*	31	15	15	91	71	185
$\beta_1 = 0.10$	$n_\tau - n^*$	66	8	-2	4	-3	5
$\beta_0 = 0.05$	n^*	153	63	60	379	303	779
$\beta_1 = 0.05$	$n_\tau - n^*$	254	17	-10	20	-7	19
$\beta_0 = 0.01$	n^*	4 877	1 591	1 495	9 633	7 767	19 751
$\beta_1 = 0.01$	$n_\tau - n^*$	5 451	435	-94	493	-153	500

лемме 1.3, асимптотическое приближение n_τ необходимого объёма выборки в этом случае совпадает с целой частью $\lceil z \rceil$ решения уравнения

$$\int_0^\infty \Phi\left(\frac{t(z) - \theta}{\sigma/\sqrt{z}}\right) \frac{1}{\tau} \phi\left(\frac{\theta - (\mu - \theta_0)}{\tau}\right) d\theta = \beta_0 \frac{\Pi_0 - \beta_1}{1 - \beta_0 - \beta_1},$$

$$t(z) = \mu - \theta_0 + \sqrt{\tau^2 + \frac{\sigma^2}{z}} \Phi^{-1}\left(\frac{\Pi_0 - \beta_1}{1 - \beta_0 - \beta_1}\right), \quad z > 0,$$

где для распределения Бернулли $\sigma^2 = 1/I_0 = \theta_0(1 - \theta_0)$. Априорная вероятность справедливости нулевой гипотезы

$$\Pi_0 = \mathbf{P}(\vartheta \leq \theta_0) = \mathbb{B}(\theta_0; p, q) \approx \Phi\left(\frac{\theta_0 - \mu}{\tau}\right) = \Phi(-\mu_0).$$

В случае $\mu = \theta_0$ имеем $\tau^2 = pq/(p + q)^3 = \mu(1 - \mu)/(p + q) = \theta_0(1 - \theta_0)/(p + q)$ и $1/(I_0\tau^2) = p + q$. Поэтому асимптотическая формула (1.46) для $\beta_0 = \beta_1 = \beta$ принимает вид

$$n_\tau = (p + q) \operatorname{ctg}^2(\pi\beta).$$

Критерий принимает гипотезу H_0 , если для $n = n_\tau$

$$\bar{x} \leq \theta_0 + t(\sqrt{n}) \tag{1.49}$$

Замечание 10. В первом столбце таблицы 1.5 модельные параметры типичны для задач контроля качества. Граничное значение $\theta_0 = 0.01$ параметра соответствует допустимой доле дефектных единиц в контролируемой партии продукции. Если уровень входного качества производимой продукции достаточно высок, то значение параметра p априорного распределения будет достаточно малым, а значение параметра q — большим. Для иллюстрации были

выбраны значения $p = 0.1$, $q = 10$, при которых уровень входного качества $Q_{in} = \mathbf{P}\{\vartheta \leq 0.01\} = 0.824$ не столь велик, чтобы можно было вовсе отказаться от процедуры контроля. Несмотря на то что $\mathbf{E}\vartheta = 1/101$ близко к границе $\theta_0 = 0.01$ и параметр концентрации априорного распределения $\mathbf{D}\vartheta = 0.000883$ весьма мал, схема стягивающегося априори практически не работает. Это объясняется существенным отличием априорного распределения от нормального. В остальных столбцах этой таблицы значения модельных параметров обеспечивают хорошую аппроксимацию априорного распределения нормальным законом, и поэтому здесь при умеренных ограничениях β_0, β_1 схема стягивающегося априори даёт результаты, не столь далёкие от истины (не более 30%).

1.3 Последовательные d-гарантийные критерии

В этом параграфе предлагаются несколько последовательных схем для построения d-гарантийных процедур, удовлетворяющих заданным ограничениям на обе d-апостериорные ошибки.

1.3.1 Процедура первого перескока

В d-апостериорном подходе, в отличие от классического, существует универсальный способ построения гарантийной статистической процедуры для любой проблемы статистического вывода (см. [7]). В проблеме различения двух гипотез гарантийный статистический вывод может быть осуществлён с помощью следующей процедуры первого перескока. Перепишем сначала формулу для d-риска, воспользовавшись (1.4) и теоремой Фубини (справедливой здесь, т.к. подынтегральные функции неотрицательны):

$$\begin{aligned} \mathcal{R}_1(\langle \nu, \delta \rangle) &= \\ &= \frac{1}{\Psi(d_1)} \sum_{n=1}^{\infty} \int_{\Theta_0} \int_{\mathbb{X}^n} \varphi_a(x^{(n)}) \mathbf{P}(\nu = n | x^{(n)}) f_n(x^{(n)} | \theta) g(\theta) \mu_n(dx^{(n)}) \gamma(d\theta) = \\ &= \frac{1}{\Psi(d_1)} \sum_{n=1}^{\infty} \int_{\mathbb{X}^n} \pi_0(x^{(n)}) \varphi_a(x^{(n)}) \mathbf{P}(\nu = n | x^{(n)}) f_n(x^{(n)}) \mu_n(dx^{(n)}), \quad (1.50) \end{aligned}$$

где

$$\Pi_0(x^{(n)}) = \mathbf{P}\{\vartheta \in \Theta_0 \mid x^{(n)}\} = \frac{1}{f_n(x^{(n)})} \int_{\Theta_0} f_n(x^{(n)} \mid \theta) g(\theta) \gamma(d\theta) \quad (1.51)$$

есть апостериорная вероятность справедливости нулевой гипотезы. Представление (1.50) можно формально записать в следующей компактной форме через оператор условного математического ожидания:

$$\mathcal{R}_1(\langle \nu, \delta \rangle) = \mathbf{E}[\Pi_0(X^{(\nu)}) \mid \delta = d_1].$$

Аналогично,

$$\mathcal{R}_0(\langle \nu, \delta \rangle) = \mathbf{E}[(1 - \Pi_0(X^{(\nu)})) \mid \delta = d_0].$$

Таким образом, если решение $\delta = d_1$ принимается только тогда, когда апостериорная вероятность $\Pi_0(X^{(\nu)}) \leq \beta_1$, а решение $\delta = d_0$ — когда $\Pi_0(X^{(\nu)}) \geq 1 - \beta_0$, то критерий $\langle \nu, \delta \rangle$ будет удовлетворять заданным ограничениям на обе d-апостериорные вероятности ошибок. Исходя из этих соображений, в [7] был предложен следующий последовательный критерий, названный процедурой первого перескока. На каждом шаге $n \geq 1$ область продолжения эксперимента определяется как $O_n = \{x^{(n)} \in \mathbb{X}^n : \beta_1 < \Pi_0(x^{(n)}) < 1 - \beta_0\}$, т.е. процесс накопления наблюдений завершается на шаге n , при котором впервые апостериорная вероятность $\Pi_0(x^{(n)})$ выходит из интервала $(\beta_1, 1 - \beta_0)$. Если остановка произошла при выходе за верхнюю границу интервала, т.е. $\Pi_0(x^{(n)}) \geq 1 - \beta_0$, то принимается нулевая гипотеза; при выходе за нижнюю границу принимается альтернативное утверждение. Момент остановки, соответствующий такому правилу продолжения наблюдений, будем обозначать ν_{un} (от английского universal procedure).

Для многих вероятностных моделей наблюдений момент остановки процедуры первого перескока может быть описан как момент первого выхода случайной суммы $S_n = \sum_1^n X_i$, $n = 1, 2, \dots$, за двусторонние границы параболического типа. Справедлива следующая

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

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

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

Доказательство. Если $\nu = \infty$, то для соответствующей последовательности реализаций $\mathbf{x} = (x_1, x_2, \dots)$ при любом $n \geq 1$ выполняются неравенства

$$\frac{a_{0n} - n\mu}{\sigma\sqrt{n \ln(\ln n)}} < \sum_{i=1}^n \frac{(x_i - \mu)}{\sigma\sqrt{n \ln(\ln n)}} < \frac{a_{1n} - n\mu}{\sigma\sqrt{n \ln(\ln n)}}.$$

По условию леммы левая и правая части этих неравенств имеют одинаковые предельные значения: 0 при $\mu = 0$, $-\infty$ при $\mu > 0$ и $+\infty$ при $\mu < 0$. По закону повторного логарифма множество таких реализаций \mathbf{x} имеет меру нуль. \square

Для рассмотренных далее вероятностных моделей момент остановки ν_{un} замкнут и, кроме того, обладает большой вероятностью остановки на первых шагах обследования, но, по-видимому, имеет бесконечное среднее значение.

Процедура первого перескока совпадает с последовательной процедурой вальдовского типа для зависимых наблюдений, область продолжения наблюдений которой определяется неравенствами

$$B < \frac{p_G(x^{(n)} | \Theta_1)}{p_G(x^{(n)} | \Theta_0)} < A, \quad (1.52)$$

где

$$p_G(x^{(n)} | \Theta_k) = \frac{1}{\Pi_k} \int_{\Theta_k} f_n(x^{(n)} | \theta) dG(\theta) \quad (1.53)$$

плотность условного (маргинального) распределения выборки относительно события $\vartheta \in \Theta_k$, $k = 0, 1$, границы $B < A$ однозначно определяются через ограничения на d-риски (см. (1.56)).

Последовательный критерий, использующий взвешенное правдоподобие, изучался ещё А. Вальдом [5, Раздел 4.1.3] (см. также [40]).

Определение 1.2. Последовательный критерий отношения вероятностей (ПКОВ) с областью продолжения наблюдений, определяемой неравенствами (1.52), где границы $B = \alpha_1/(1 - \alpha_0)$, $A = (1 - \alpha_1)/\alpha_0$, назовём критерием вальдовского типа уровня (α_0, α_1) .

Замечание 11. В случае, когда априорное распределение G сосредоточено в двух точках $\theta_0 \in \Theta_0$, $\theta_1 \in \Theta_1$, классические вероятности ошибок первого и второго рода критерия вальдовского типа уровня (α_0, α_1) приблизительно совпадают с (α_0, α_1) (см. [5]).

Лемма 1.12. Пусть $\Pi_k = \mathbf{P}(\vartheta \in \Theta_k)$, $k = 0, 1$, — априорные вероятности проверяемых гипотез.

1. Если $\beta_1 < \Pi_0 < 1 - \beta_0$, то последовательная процедура первого перескока есть критерий вальдовского типа уровня (α_0, α_1) с

$$\alpha_0 = \frac{\beta_1(\Pi_1 - \beta_0)}{\Pi_0(1 - \beta_0 - \beta_1)}, \quad \alpha_1 = \frac{\beta_0(\Pi_0 - \beta_1)}{\Pi_1(1 - \beta_0 - \beta_1)} \quad (1.54)$$

2. Если априорные вероятности гипотез совпадают ($\Pi_0 = \Pi_1 = 1/2$) и ограничения на d -риски также равны между собой ($\beta_0 = \beta_1 = \beta$), то последовательная процедура первого перескока есть критерий вальдовского типа уровня (β, β) .

Доказательство. Легко видеть, что апостериорную вероятность (1.51) можно представить в виде

$$\Pi_0(x^{(n)}) = \left(1 + \frac{p_G(x^{(n)} | \Theta_1) \Pi_1}{p_G(x^{(n)} | \Theta_0) \Pi_0} \right)^{-1}. \quad (1.55)$$

С учётом этого представления область продолжения процедуры первого перескока может быть записана в виде (1.52) с границами

$$B = \frac{\beta_0}{1 - \beta_0} \frac{\Pi_0}{\Pi_1}, \quad A = \frac{1 - \beta_1}{\beta_1} \frac{\Pi_0}{\Pi_1}. \quad (1.56)$$

Отсюда сразу следует часть 2 леммы. Для доказательства первой части достаточно заметить, что заявленные в (1.54) значения α_0 и α_1 суть решения системы уравнений $B = \alpha_1/(1 - \alpha_0)$, $A = (1 - \alpha_1)/\alpha_0$. \square

1.3.2 Усечённая процедура первого перескока

Зафиксируем n и предположим, что правило принятия решений φ_a для этого n таково, что

$$\int_{\mathbb{X}^n} \pi_0(x^{(n)}) \varphi_a(x^{(n)}) f_n(x^{(n)}) \mu_n(dx^{(n)}) \leq \beta_1 \Psi_n(d_1),$$

где $\Psi_n(d_1) = \int_{\mathbb{X}^n} \varphi_a(x^{(n)}) f_n(x^{(n)}) \mu_n(dx^{(n)})$. Другими словами, процедура с фиксированным числом наблюдений $\langle n, \varphi_a \rangle$ удовлетворяет заданному ограничению β_1 на величину d-риска $\mathcal{R}_1(n, \varphi_a)$. Тогда соответствующее этому n слагаемое в формуле для d-риска (1.50)

$$\int_{\mathbb{X}^n} \pi_0(x^{(n)}) \varphi_a(x^{(n)}) \mathbb{P}(\nu = n | x^{(n)}) f_n(x^{(n)} | \theta) \mu_n(dx^{(n)}) \leq \beta_1 \Psi_n(d_1).$$

Рассмотрим усечённый вариант универсальной последовательной процедуры $\langle \tilde{\nu}^*, \varphi_a \rangle$, который в отличие от последней останавливается с вероятностью единица при объёме выборки, равном НОВ n^* . Правила принятия решений для $n < n^*$ имеют вид

$$\varphi_a(x^{(n)}) = \mathbb{I}\{\pi_0(x^{(n)}) \leq \beta_1\},$$

а для $n = n^*$ правило φ_a совпадает с правилом (1.6), для которого справедливы соотношения (1.7)–(1.8). Следовательно, d-риск усечённой универсальной последовательной процедуры

$$\begin{aligned} \mathcal{R}_1(\langle \tilde{\nu}^*, \varphi_a \rangle) &\leq \\ &\leq \beta_1 \frac{\sum_1^{n^*-1} \int_{\mathbb{X}^n} \varphi_a(x^{(n)}) \mathbb{P}(\tilde{\nu}^* = n | x^{(n)}) f_n(x^{(n)}) \mu(dx^{(n)}) + \Psi_{n^*}(d_1)}{\sum_1^{n^*} \int_{\mathbb{X}^n} \varphi_a(x^{(n)}) \mathbb{P}(\tilde{\nu}^* = n | x^{(n)}) f_n(x^{(n)}) \mu(dx^{(n)})}. \end{aligned} \quad (1.57)$$

Ясно, что $\Psi_{n^*}(d_1) > \int_{\mathbb{X}^{n^*}} \varphi_a(x^{(n^*)}) \mathbb{P}(\tilde{\nu}^* = n^* | x^{(n^*)}) f_{n^*}(x^{(n^*)}) \mu(dx^{(n^*)})$. Поэтому правая часть неравенства (1.57) заведомо больше β_1 . Заметим, однако, что при доказательстве гарантийности универсальной последовательной процедуры использовалось неравенство $\Pi_0(x^{(n)}) \leq \beta_1$, которое в действительности является строгим. Поэтому можно ожидать, что это компенсирует превышение над β_1 в неравенстве (1.57) или же это превышение будет не существенным (см. примеры ниже).

Аналогичные выводы справедливы и для d-риска $\mathcal{R}_0(\langle \tilde{\nu}^*, \varphi_a \rangle)$ усечённой процедуры. Заметим ещё, что если по условиям задачи строго необходимо, чтобы один из рисков, например \mathcal{R}_0 , не превосходил заданного уровня β_0 , то можно в усечённой последовательной процедуре в случае, когда $\tilde{\nu}^* = n^*$, принимать решение в пользу нулевой гипотезы только при $\Pi_0(\mathbf{x}^{(n^*)}) \geq 1 - \beta_0$. Для такого варианта процедуры d-риск $\mathcal{R}_0(\langle \tilde{\nu}^*, \delta \rangle) \leq \beta_0$, при этом увеличение d-риска $\mathcal{R}_1(\langle \tilde{\nu}^*, \delta \rangle)$ по сравнению с β_1 , к сожалению, может оказаться значительным.

1.3.3 Последовательный критерий на статистике вклада

Ещё одна последовательная процедура, характеристики которой анализируются в данной работе, основана на статистике вклада

$$T_n(X^{(n)}) = \sum_{i=1}^n \left. \frac{\partial \ln f(X_i | \theta)}{\partial \theta} \right|_{\theta=\theta_0}.$$

Будем называть эту процедуру последовательным критерием вклада, а соответствующий ей момент останковки обозначать ν_{sc} (от английского score function). Область продолжения наблюдений задаётся неравенствами

$$-A < T_n(x^{(n)}) < B,$$

с границами, определяемыми из системы уравнений ($r > 0$, $0 < p < 1$)

$$A = (1-p)r, \quad B = pr, \tag{1.58}$$

$$\frac{\mathbf{C} + \Psi(1-p)}{\beta_0 \Pi_0} = \frac{\mathbf{C} + \Psi(p)}{\beta_1 \Pi_1}, \tag{1.59}$$

$$r = \frac{g(\theta_0)(1-\beta_0-\beta_1)(-\mathbf{C}-\Psi(p))}{2\beta_1 \Pi_1},$$

где $\Psi(p) = \Gamma'(p)/\Gamma(p)$ — дигамма-функция Эйлера, константа $\mathbf{C} = -\Psi(1) \approx \approx 0.577216$ есть константа Эйлера.

В соответствии с леммой 1.11 для регулярных экспериментов с конечной информацией Фишера $I(\theta) = \mathbf{E}_\theta(\partial \ln f(X | \theta)/\partial \theta)^2 \neq 0$ момент останковки ν_{sc} замкнут. Кроме того, хорошо известно (см., например, [5]), что в этом случае $\mathbf{E}_\theta \nu_{sc} < \infty$ при любом $\theta \in \Theta$.

Асимптотическое поведение среднего значения момента останковки ν_{sc} , когда $\beta_0 + \beta_1 \rightarrow 0$ и при этом $\theta \rightarrow \theta_0$ подробно изучено в статье [58]. В частности, в этой статье показано, что если $\beta = \beta_0 = \beta_1 \rightarrow 0$, то безусловное среднее $\mathbf{E}[\nu_{sc}] = O(\beta^{-1} \ln \beta)$ и что для d-гарантийного последовательного критерия, основанного на статистике вклада, математическое ожидание $\mathbf{E}_\theta[\nu_{sc}]$ меньше НОВ при всех θ . Для сравнения отметим, что последовательный критерий Вальда, гарантирующий заданные вероятности классических ошибок первого и второго рода, при некоторых θ может иметь $\mathbf{E}_\theta[\nu]$ большее НОВ.

Последовательные процедуры в модели N–N. Рассмотрим задачу проверки гипотезы $H_0 : \theta \leq 0$ против альтернативы $H_1 : \theta > 0$ в рамках модели N–N (1.10). Так как апостериорное распределение ϑ при фиксированном результате наблюдений $x^{(n)}$ нормально с параметрами (1.11), то апостериорная вероятность области $\Theta_0 = (-\infty, 0]$ равна $\Pi_0(x^{(n)}) = \Phi(-m_n/\tau_n)$. Таким образом, область продолжения наблюдений универсальной гарантийной процедуры (процедуры первого перескока) определяется неравенствами

$$A_{0n} < \sum_{i=1}^n X_i < A_{1n}$$

с границами параболического типа:

$$\begin{aligned} A_{0n} &= -\mu \frac{\sigma^2}{\tau^2} - \sigma q_0 \sqrt{n + \frac{\sigma^2}{\tau^2}}, \\ A_{1n} &= -\mu \frac{\sigma^2}{\tau^2} + \sigma q_1 \sqrt{n + \frac{\sigma^2}{\tau^2}}, \end{aligned} \tag{1.60}$$

где $q_k = \Phi^{-1}(1 - \beta_k)$, $k = 0, 1$. В силу леммы 1.11 эта процедура останавливается с вероятностью единица при любом фиксированном $\theta \in \mathbb{R}^1$: $\mathbf{P}_\theta(\nu_{un} < \infty) = 1$. Следовательно, и относительно безусловного распределения $\mathbf{P}(\nu_{un} < \infty) = 1$.

Покажем, что при любом $\theta \neq 0$ среднее значение момента остановки ν_{un} конечно.

Теорема 1.5. Пусть X_1, X_2, \dots — независимые нормальные $\mathcal{N}(\theta, \sigma^2)$ случайные величины. Тогда для любого $\theta \neq 0$ среднее значение момента остановки ν_{un} процедуры первого перескока $\mathbf{E}_\theta[\nu_{un}] < \infty$.

Доказательство. Упрощая изложение, будем считать $\mu = 0$ и $\sigma^2 = \tau^2 = 1$. Рассмотрим только случай $\theta > 0$ (случай $\theta < 0$ доказывается аналогично).

Так как сумма n независимых нормальных $\mathcal{N}(\theta, \sigma^2)$ случайных величин имеет нормальное $\mathcal{N}(n\theta, n\sigma^2)$ распределение, то для любого $n \geq 1$

$$\mathbf{P}_\theta(\nu_{un} > n) \leq \mathbf{P}_\theta\left(A_{0n} < \sum_{i=1}^n X_i < A_{1n}\right) =$$

$$\begin{aligned}
&= \Phi\left(\frac{A_{1n} - n\theta}{\sqrt{n}\sigma}\right) - \Phi\left(\frac{A_{0n} - n\theta}{\sqrt{n}\sigma}\right) = \\
&= \phi(\tilde{A}_n) \sigma \frac{(q_1 + q_0)}{\sqrt{n}} \sqrt{n + \frac{\sigma^2}{\tau^2}}, \tag{1.61}
\end{aligned}$$

где последнее равенство следует из формулы конечных приращений с соответствующей «средней точкой» $\tilde{A}_n \in ((A_{0n} - n\theta)/(\sqrt{n}\sigma), (A_{1n} - n\theta)/(\sqrt{n}\sigma))$. Следовательно, средняя точка в (1.61) удовлетворяет неравенствам

$$-q_0 \sqrt{1 + \frac{1}{n}} - \sqrt{n}\theta < \tilde{A}_n < q_1 \sqrt{1 + \frac{1}{n}} - \sqrt{n}\theta.$$

Для любого фиксированного $\theta > 0$, начиная с некоторого n_0 обе части этого соотношения отрицательны при всех $n \geq n_0$. Поэтому

$$\begin{aligned}
\phi(\tilde{A}_n) &< \phi\left(q_1 \sqrt{1 + n^{-1}} - \sqrt{n}\theta\right) = \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(q_1 \sqrt{1 + n^{-1}} - \sqrt{n}\theta\right)^2\right) \leq \\
&\leq \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(n\theta^2 - 2q_1\theta\sqrt{n+1})\right) \leq \\
&\leq \frac{1}{\sqrt{2\pi}} \int_{n+1}^{n+2} e^{-\theta^2(x-2)/2} e^{q_1\theta\sqrt{x}} dx,
\end{aligned}$$

где справедливость последнего неравенства следует из того, что функция $e^{q_1\theta\sqrt{x}}$ возрастает по x , а функция $e^{-\theta^2(x-2)/2}$ убывает по x . Так как, очевидно, что $\sqrt{n+1}/\sqrt{2\pi n} \leq 1$ при всех $n \geq 1$, то

$$\begin{aligned}
\mathbf{E}_\theta[\nu_{un}] &= \sum_{n=0}^{\infty} \mathbf{P}_\theta(\nu_{un} > n) \leq \\
&\leq n_0 - 1 + (q_0 + q_1) \sum_{n=n_0}^{\infty} \int_{n+1}^{n+2} e^{-\theta^2(x-2)/2} e^{q_1\theta\sqrt{x}} dx = \\
&= n_0 - 1 + (q_0 + q_1) e^{\theta^2} \int_{n_0+1}^{\infty} e^{-\theta^2 x/2} e^{q_1\theta\sqrt{x}} dx.
\end{aligned}$$

Произведя в интеграле замену переменной $x \rightarrow z^2/\theta^2$, $z > 0$, получаем

$$\mathbf{E}_\theta[\nu_{un}] \leq n_0 - 1 + 2(q_0 + q_1) e^{\theta^2} \frac{1}{\theta^2} \int_0^{\infty} z e^{-z^2/2} e^{q_1 z} dz < \infty,$$

что и требовалось. □

Замечание 12. Оценка для математического ожидания $\mathbf{E}_\theta[\nu_{un}]$, полученная в предыдущей теореме, не даёт оснований для утверждения о конечности безусловного среднего $\mathbf{E}[\nu_{un}]$, поскольку функция $1/\theta^2$ не интегрируема в нуле. Результаты численного моделирования показывают, что хвост $\bar{F}(n) = \mathbf{P}\{\nu_{un} > n\}$ распределения ν_{un} убывает медленнее, чем $1/n$ (см. рисунок 1.1). Точнее, зависимость $\bar{F}(n)$ от n может быть описана уравнением $\bar{F}(n) = c/n^\alpha$, $c > 0$, $0 < \alpha < 1$. Среднее значение для такого распределения не существует.

Ещё одним подтверждением нашего предположения о бесконечности безусловного среднего ν_{un} служит следующее утверждение.

Теорема 1.6. *Если $q = \min(q_1, q_2) > 1$, т.е. ограничения $\beta_0, \beta_1 \leq 0.158$, и априорное среднее $|\mu| \leq \sqrt{q^2 - 1}$, то при $\theta = 0$ среднее значение момента остановки процедуры первого перескока $\mathbf{E}_0 \nu_{un} = \infty$.*

Доказательство. Снова для упрощения выкладок будем считать, что $\sigma^2 = \tau^2 = 1$ и, кроме того, $q_0 = q_1 = q$ (т.е. ограничения $\beta_0 = \beta_1 = \beta$). Пусть $\mu \geq 0$. Предположим, что $\mathbf{E}_0 \nu_{un} < \infty$, тогда в силу тождества Вальда для суммы $S_n = \sum_1^n X_i$ независимых центрированных случайных величин (см., например, [26]) справедливы соотношения

$$\begin{aligned} \mathbf{E}_0 \nu_{un} &= \mathbf{E}_0 (S_{\nu_{un}})^2 \geq \\ &\geq \mathbf{E}_0 [\min \{ (q\sqrt{\nu_{un} + 1} + \mu)^2, (q\sqrt{\nu_{un} + 1} - \mu)^2 \}] = \\ &= \mathbf{E}_0 [(q\sqrt{\nu_{un} + 1} - \mu)^2] = \\ &= q^2 \mathbf{E}_0 (\nu_{un} + 1) - 2\mu q \mathbf{E}_0 \sqrt{\nu_{un} + 1} + \mu^2, \end{aligned}$$

т.к. на момент остановки сумма $S_{\nu_{un}} \geq -\mu + q\sqrt{\nu_{un} + 1} > 0$ или $S_{\nu_{un}} \leq -\mu - q\sqrt{\nu_{un} + 1} < 0$. Таким образом,

$$\mathbf{E}_0 [(1 - q^2)\xi^2 + 2\mu q\xi] \geq \mu^2 + 1,$$

где случайная величина $\xi = \sqrt{\nu_{un} + 1}$. Максимум квадратической функции $-(q^2 - 1)x^2 + 2\mu qx$, $x > 0$, не превосходит $\mu^2 q^2 / (q^2 - 1)$. Таким образом, имеем неравенство $\mu^2 q^2 / (q^2 - 1) > \mu^2 + 1$, которое противоречит условию теоремы $\mu^2 \leq q^2 - 1$. \square

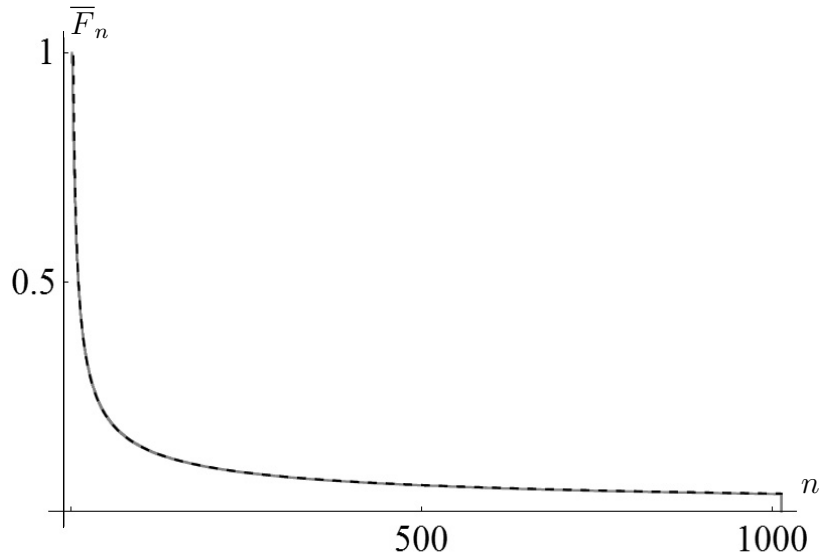


Рис. 1.1: Оценка функции надёжности $\bar{F}(n)$ момента остановки ν_{un}

Замечание 13. Момент остановки $\nu_{un} \neq 0$ только в случае, когда $\beta_0 < \Phi(\mu) < 1 - \beta_1$, т.е. при $|\mu| < \min(q_0, q_1) = q$. Для $\beta_0, \beta_1 \leq 0.05$ эта граница на μ не сильно превосходит границу из условия теоремы.

Замечание 14. В подтверждение предположения о бесконечном значении математического ожидания $\mathbf{E}[\nu_{un}]$ момента остановки универсальной процедуры на рисунке 1.1 для $\mu = 0$, $\beta_0 = \beta_1 = 0.01$ приведён график эмпирической функции надёжности $\bar{F}_k(n)$ момента остановки ν_{un} , ограниченного сверху НОВ $n^* = 1013$ (серая линия), а также график оценки функции надёжности посредством степенной функции $0.603 n^{-0.56}$ (пунктирная линия). Для построения последней эмпирическая функция надёжности $\bar{F}(n)$ приводилась к логарифмической шкале, т.е. бралась функция $y = \ln(\bar{F}(e^u))$ ($u = \ln n$), которая подгонялась линейной функцией методом наименьших квадратов. Как видно, графики на рисунке совпадают идеально.

Перейдём к анализу последовательной процедуры на статистике вклада. Статистика вклада для нормальной модели наблюдений

$$\sum_{i=1}^n \frac{\partial \ln f(x_i | \theta)}{\partial \theta} \Big|_{\theta=0} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \theta) \Big|_{\theta=0} = \frac{1}{\sigma^2} \sum_{i=1}^n x_i.$$

Поэтому последовательный критерий вклада имеет область продолжения наблюдений

$$-\sigma^2 A < S_n = \sum_1^n X_i < \sigma^2 B, \quad n \geq 1, \quad (1.62)$$

где константы A, B вычисляются по формулам (1.58)–(1.59), в которых значение функции $g(\theta_0) = \phi(\theta_0/\tau)/\tau = 1/\sqrt{2\pi}\tau$.

Эта последовательная схема хорошо изучена. В частности, момент остановки ПКСВ обладает конечным средним значением $\mathbf{E}_\theta \nu_{sc}^k$ для любого $k > 0$ и любого $\theta \in \mathbb{R}^1$, так как хвост его распределения убывает к нулю с экспоненциальной скоростью. Осталось доказать, что и по безусловному распределению математическое ожидание $\mathbf{E} \nu_{sc}$ конечно. Заметим, что относительно безусловного распределения \mathbf{P} случайные величины X_1, \dots, X_n уже не являются независимыми.

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

Доказательство. Зафиксируем некоторое $\theta \in \mathbb{R}^1$. Тогда в силу независимости и одинаковой распределенности X_1, \dots, X_n относительно распределения \mathbf{P}_θ вероятность того, что последовательная процедура не остановится до момента n

$$\begin{aligned} \mathbf{P}_\theta(\nu_{sc} > n) &= \mathbf{P} \left(\bigcap_{i=1}^n \{ -\sigma^2 A < S_n < \sigma^2 B \} \right) \leq \\ &\leq \mathbf{P}_\theta \left(\bigcap_{i=1}^n \{ |X_i| < C \} \right) = \left(\mathbf{P}_\theta(|X_1| < C) \right)^n, \end{aligned}$$

где $C = \sigma^2 A + \sigma^2 B$ (см. [5, стр. 202] или [26, стр. 100]). Так как стандартное нормальное распределение симметрично около нуля и его плотность имеет единственный максимум в нуле, то для всех $\theta \in \mathbb{R}^1$

$$\mathbf{P}_\theta(|X_1| < C) = \Phi\left(\frac{C-\theta}{\sigma}\right) - \Phi\left(\frac{-C-\theta}{\sigma}\right) \leq \Phi\left(\frac{C}{\sigma}\right) - \Phi\left(-\frac{C}{\sigma}\right) = q < 1.$$

Поэтому для любого $\theta \in \mathbb{R}^1$ вероятность $\mathbf{P}_\theta(\nu_{sc} > n) \leq q^n \xrightarrow{n \rightarrow \infty} 0$ и математическое ожидание

$$\mathbf{E}_\theta \nu_{sc} = 1 + \sum_{n=1}^{\infty} \mathbf{P}_\theta(\nu_{sc} > n) \leq 1 + \sum_{n=1}^{\infty} q^n = 1 + \frac{1}{1-q} < \infty.$$

Так как q не зависит от θ , то по безусловному распределению $\mathbf{P}(\nu_{sc} > n) = \mathbf{E} \mathbf{P}_\theta(\nu_{sc} > n) \leq q^n \rightarrow 0$. Следовательно, и безусловное среднее $\mathbf{E} \nu_{sc} \leq 1 + 1/(1-q)$. \square

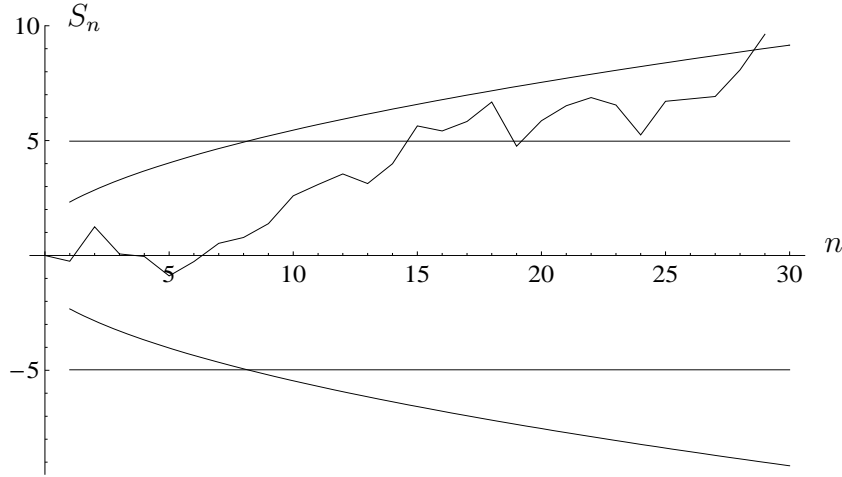


Рис. 1.2: Траектория S_n в границах ПКОВ и ПКСВ

Рисунок 1.2 иллюстрирует траекторию последовательных значений статистики $S_n = \sum_1^n X_i$, $n = 1, 2, \dots$, между границами областей продолжения наблюдений в процедурах φ_{sc} (прямые $y = \pm 4.97$) и φ_{un} (параболические границы $y = \pm 1.64\sqrt{n+1}$) при $\mu = 0$, $\tau = \sigma = 1$, $\beta_1 = \beta_2 = 0.05$ и случайно сгенерированном значении $\theta = 0.505$.

На рисунке хорошо видно, что процедура первого перескока обладает значительно большей вероятностью остановки на первых шагах эксперимента. Заметим, что при указанных значениях параметров модели НОВ с фиксированным числом наблюдений $n^* = 40$.

Последовательные процедуры для модели G-E. Изучим теперь последовательные процедуры различения гипотез $H_0 : \theta \leq 1$ и $H_1 : \theta > 1$ в рамках гамма-показательной модели (1.18).

В силу утверждений леммы 1.4 апостериорная вероятность $\mathbf{P}(\vartheta \leq 1 | \bar{x}) = \mathbb{G}(a + n\bar{x}, n + \lambda, 1)$. Поэтому область продолжения наблюдений универсальной гарантийной процедуры

$$\mathbb{G}^{-1}(\beta_1; n + \lambda, 1) - a < \sum_1^n X_i < \mathbb{G}^{-1}(1 - \beta_0; n + \lambda, 1) - a.$$

Решение в пользу гипотезы $H_0 : \theta \leq \theta_0$ ($= 1$) принимается в том случае, когда последовательный процесс заканчивается выходом за правую границу. Из (1.21) следует справедливость следующего утверждения.

Лемма 1.13. I) Если ϑ имеет гамма-распределение $\mathcal{G}(\lambda, a)$, а наблюдения X_i при фиксированном $\vartheta = \theta$ имеют показательное распределение со средним $1/\theta$, то область продолжения наблюдений процедуры первого перескока определяется неравенствами

$$\begin{aligned} -\sqrt{n+\lambda}q_1 + A_{1n} &< \sum_{i=1}^n (x_i - 1) < \sqrt{n+\lambda}q_0 + A_{0n}, \\ A_{1n} &= \lambda - a + \frac{1}{3}(q_1^2 - 1) + C_{1n}, \\ A_{0n} &= \lambda - a + \frac{1}{3}(q_0^2 - 1) + C_{0n}, \end{aligned} \tag{1.63}$$

где $q_j = \Phi^{-1}(1-\beta_j)$, $j = 0, 1$, и $C_{0n}, C_{1n} \rightarrow 0$ ($n \rightarrow \infty$), причём $|C_{0n}|, |C_{1n}| \leq 1$ при всех $n \geq 1$ для $\beta_0, \beta_1 > 0.001$.

II) Момент остановки процедуры первого перескока для гамма-показательной модели замкнут, т.е. $\mathbf{P}(\nu_{un} < \infty) = 1$ при $\forall \theta \in [0, 1]$.

Таким образом, как и в случае модели N-N, процедура первого перескока может быть описана как случайное блуждание накопленной суммы $\sum_1^n (X_i - 1)$ между двумя параболическими границами. Поэтому здесь также справедлив факт о замкнутости момента остановки ν_{un} и, скорее всего, о бесконечности его безусловного среднего значения.

Статистика вклада для показательной модели наблюдений

$$\sum_{i=1}^n \left. \frac{\partial \ln f(x_i | \theta)}{\partial \theta} \right|_{\theta=1} = \sum_{i=1}^n \left(-x_i + \frac{1}{\theta} \right) \Big|_{\theta=1} = -\sum_{i=1}^n (x_i - 1).$$

Поэтому последовательный критерий вклада продолжает наблюдения до тех пор, пока

$$-B < \sum_1^n (x_i - 1) < A, \tag{1.64}$$

где константы A, B вычисляются по формулам (1.58)–(1.59). Выход за правую границу приводит к принятию гипотезы $H_0 : \theta \leq 1$, за левую — гипотезы $H_1 : \theta > 1$.

Последовательные процедуры для модели В-В. Изучим свойства последовательных d-гарантийных процедур различения двух гипотез $H_0 :$

$\theta \leq \theta_0$ и $H_1 : \theta > \theta_0$, где фиксированное значение $\theta_0 \in (0, 1)$, в рамках модели бета-Бернулли (1.23).

Так как апостериорная вероятность нулевой гипотезы при фиксированном значении статистики S_n равна $\mathbb{B}(\theta_0; p + S_n, q + n - S_n)$ (см. (1.25)), то процедура первого перескока определяется областью продолжения наблюдений

$$\beta_1 < \mathbb{B}(\theta_0; p + S_n, q + n - S_n) < 1 - \beta_0. \quad (1.65)$$

Гипотеза H_0 принимается, когда апостериорная вероятность нулевой гипотезы становится больше $1 - \beta_0$.

Как и для двух предыдущих моделей, здесь границы продолжения наблюдений также имеют параболический тип. Этот факт основан на следующем предложении, которое может быть выведено из представления Уишарта для функции бета-распределения (см., например, описание к таблицам функции бета-распределения в [3]).

Лемма 1.14. Пусть $x_n = x_n(\beta)$ есть решение уравнения $\mathbb{B}(\theta_0; p + x_n, q + n - x_n) = \beta$. Тогда для любых $p, q > 0$, $\theta_0, \beta \in (0, 1)$ справедливо асимптотическое представление

$$x_n(\beta) = n\theta_0 + \Phi^{-1}(1 - \beta) \sqrt{\theta_0(1 - \theta_0)} \sqrt{n + p + q} + \Delta_n,$$

где $\overline{\lim} |\Delta_n| < 1$.

Таким образом, справедлива

Лемма 1.15. I) Если ϑ имеет бета-распределение $\mathcal{B}(a, b)$, а наблюдаемая в эксперименте (с $\vartheta = \theta$) случайная величина X имеет распределение Бернулли со средним значением θ , то область продолжения наблюдений процедуры первого перескока определяется неравенствами

$$-q_0\sigma_0\sqrt{n + a + b} + A_{0n} < \sum_{i=1}^n (x_i - \theta_0) < q_1\sigma_0\sqrt{n + a + b} + A_{1n},$$

где $\sigma_0 = \sqrt{\theta_0(1 - \theta_0)}$ и $A_{jn} = O(1)$, $j = 0, 1$, при $n \rightarrow \infty$.

II) Момент останова процедуры первого перескока для модели В-В замкнут, т.е. $\mathbf{P}_\theta(\nu_{un} < \infty) = 1$ при любом $\theta \in [0, 1]$.

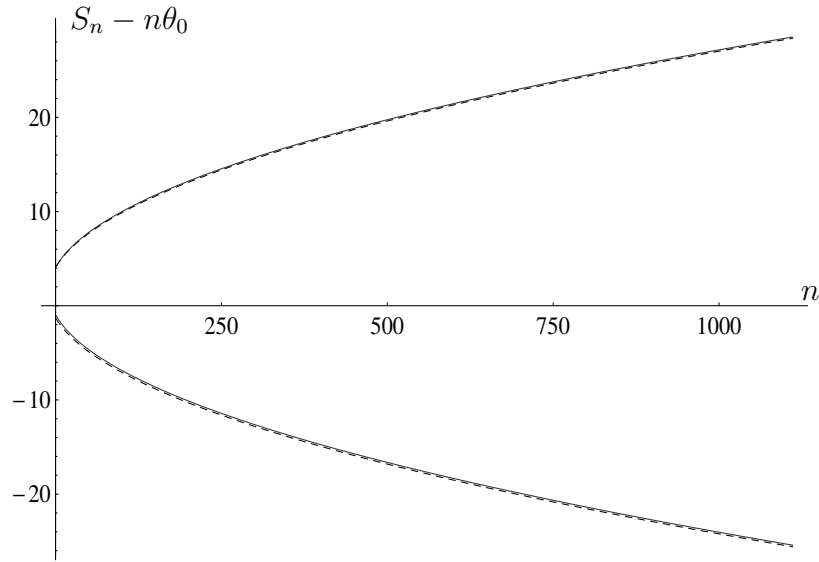


Рис. 1.3: Асимптотические и точные границы случайного блуждания ν_{un} ; В-В модель

На рисунке 1.3 приведены графики изменения границ случайного блуждания (с непрерывным параметром n), найденные по точным формулам (1.65) (сплошная линия), и их приближение из леммы 1.15 (пунктирная линия). Рассмотрен случай гипотезы $H_0 : \theta \leq 0.4$ с ограничениями $\beta_0 = \beta_1 = 0.05$ и параметрами априорного распределения $p = 3, q = 8$.

Так как статистика вклада для модели Бернулли

$$\left. \frac{\partial}{\partial \theta} \sum_{i=1}^n (x_i \ln \theta + (1 - x_i) \ln(1 - \theta)) \right|_{\theta=\theta_0} = \frac{1}{\theta_0(1 - \theta_0)} \sum_{i=1}^n (x_i - \theta_0),$$

то последовательный критерий вклада продолжает наблюдения до тех пор, пока

$$-A\theta_0(1 - \theta_0) < \sum_1^n (x_i - \theta_0) < B\theta_0(1 - \theta_0), \quad (1.66)$$

где константы A, B вычисляются по формулам (1.58), (1.59). Выход за левую границу приводит к принятию гипотезы $H_0 : \theta \leq \theta_0$, за правую — гипотезы $H_1 : \theta > \theta_0$.

1.4 Эмпирические аналоги d-гарантийных процедур

Основная проблема применения процедур в байесовском подходе состоит в правильной спецификации априорного распределения выводного параметра. Здесь необходимо отметить, что мы рассматриваем только такие ситуации,

когда имеется реальная последовательность статистических экспериментов, в каждом из которых необходимо принять решение относительно распределения наблюдаемой случайной величины. В этом случае можно предполагать, что мы обладаем не только наблюдениями в текущем эксперименте, но и во всех предыдущих экспериментах (назовём их архивными наблюдениями). Это позволяет надеяться на возможность построения состоятельной оценки истинного априорного распределения.

1.4.1 Эмпирические оценки параметров вероятностной модели

Очень часто можно предполагать, что априорное распределение G искомого параметра θ принадлежит некоторому параметрическому классу распределений: $G \in \{G_\lambda, \lambda \in \Lambda\}$. Пусть в эксперименте наблюдается n независимых случайных величин X_1, \dots, X_n , общее распределение которых описывается плотностью (относительно некоторой меры μ) $f(x | \theta, \zeta)$, где θ — параметр, о котором необходимо принять решение, ζ — мешающий параметр распределения, который предполагается одинаковым во всех экспериментах. Если значение θ есть реализация случайной величины с распределением G_λ , то последовательность k экспериментов представляет собой k независимых реализаций n -мерных случайных векторов $\mathbf{X} = (X_1, \dots, X_n)$ с плотностью

$$f_{\lambda, \zeta}(x_1, \dots, x_n) = \int_{\Theta} \prod_{i=1}^n f(x_i | \theta, \zeta) dG_\lambda(\theta).$$

Таким образом, мы имеем классическую задачу оценивания по ряду независимых наблюдений $\mathbf{x}_1, \dots, \mathbf{x}_k$. Заметим сначала, что часто мешающий параметр ζ (или некоторая непрерывная функция от него) допускает несмещенную оценку в рамках одного эксперимента, т.е. найдется статистика $Z = Z(X_1, \dots, X_n)$ такая, что $\mathbf{E}_{\theta, \zeta} Z = \zeta$ при любых θ, ζ . Очевидно, оценка $\tilde{\zeta}_k = \sum_1^k Z_j / k$, где Z_j — значение Z в j -м эксперименте, будет не только несмещенной, но в силу закона больших чисел и состоятельной (при $k \rightarrow \infty$) оценкой ζ .

В связи с необходимостью формирования архива наблюдений важно изучить вопрос о наборе статистик, которыми при этом достаточно ограничиться.

Предположим, что плотность $f(x | \theta, \zeta)$ наблюдаемой в эксперименте случайной величины принадлежит к экспоненциальному типу:

$$f(x | \theta, \zeta) = \exp \left\{ Q(\theta, \zeta)h(x) + \sum_{s=0}^m \zeta_s h_s(x) + C(\theta, \zeta) \right\},$$

где $\theta \in \Theta \subset \mathbb{R}^1$, $\zeta = (\zeta_1, \dots, \zeta_m) \in \mathbb{R}^m$, $\zeta_0 = 1$. Пусть $\mathbf{x}_j = (x_{j1}, \dots, x_{jn})$ — вектор наблюдений в j -м эксперименте, $j = 1, \dots, k$. Тогда правдоподобие относительно всей матрицы данных вычисляется как

$$\begin{aligned} & \prod_{j=1}^k f_{\lambda, \zeta}(\mathbf{x}_j) = \\ & = \exp \left\{ kn \sum_{s=0}^m \zeta_s \bar{h}_s \right\} \prod_{j=1}^k \int_{\Theta} \exp \left\{ nQ(\theta, \zeta) \bar{h}(\mathbf{x}_j) + nC(\theta, \zeta) \right\} dG_{\lambda}(\theta), \end{aligned}$$

где $\bar{h}_s = \sum_{j=1}^k \sum_{i=1}^n h_s(x_{ji})/kn$, $\bar{h}(\mathbf{x}_j) = \sum_{i=1}^n h(x_{ji})/n$. Таким образом, в общем случае достаточная статистика состоит из m статистик \bar{h}_s , $s = \overline{1, m}$, и k статистик $\bar{h}(\mathbf{X}_j)$, $j = \overline{1, k}$. Имея в архиве значения всех этих статистик, можно построить различные оценки параметров модели $(\zeta_1, \dots, \zeta_m, \lambda)$, например оценки максимального правдоподобия или оценки по методу моментов.

Оценки параметров модели N–N. Рассмотрим задачу оценивания параметров нормально-нормальной (N–N) модели (1.10), когда наблюдения $X^{(n)}$ подчинены нормальному $\mathcal{N}(\theta, \sigma^2)$ закону, а параметр θ есть реализация нормальной $\mathcal{N}(\mu, \tau^2)$ случайной величины. Модель содержит три неизвестных параметра (μ, τ^2, σ^2) , значения которых можно оценить по накопленному архиву данных. Эта модель уникальна в том смысле, что нормальность безусловного распределения \bar{X} может зачастую обеспечить нормальность условного распределения \bar{X} (при фиксированном среднем) и нормальность априорного распределения ϑ . Точнее, по теореме Г. Крамера, если плотность условного распределения наблюдений зависит от неизвестного параметра θ как $f(x - \theta)$, то безусловное распределение X нормально в том и только в том случае, когда f — плотность нормального распределения и априорное распределение ϑ также нормально.

Несмотря на такое замечательное качество N–N модель не идентифицируется полностью, если после каждого эксперимента заносить в архив только

значение \bar{x} — необходимо также фиксировать значение выборочной дисперсии s^2 . Точнее, найдем достаточную статистику последовательности наблюдений в k экспериментах $\mathbf{x}^{(k)} = (x_{ji}, i = 1, \dots, n_j, j = 1, \dots, k)$ в предположениях N–N модели.

Для выборки $x_j^{(n_j)} = (x_{j1}, \dots, x_{jn_j})$ из j -го эксперимента функция правдоподобия вычисляется как

$$L(\mu, \tau^2, \sigma^2 | x_j^{(n_j)}) = \frac{1}{(\sqrt{2\pi\sigma^2})^{n_j}} \int_{-\infty}^{\infty} \left[\prod_{i=1}^{n_j} e^{-\frac{1}{2\sigma^2}(x_{ji}-\theta)^2} \right] \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}(\theta-\mu)^2} d\theta.$$

После преобразования $\sum_1^{n_j} (x_{ji} - \theta)^2 = \sum_1^{n_j} (x_{ji} - \bar{x}_j)^2 + n_j(\bar{x}_j - \theta)^2$, где $\bar{x}_j = \sum_1^{n_j} x_{ji}/n_j$, и замены $SS_j = \sum_1^{n_j} (x_{ji} - \bar{x}_j)^2$ функция правдоподобия записывается в виде

$$L(\mu, \tau^2, \sigma^2 | x_j^{(n_j)}) = \frac{e^{-SS_j/2\sigma^2}}{\sqrt{n_j}(\sqrt{2\pi\sigma^2})^{n_j-1}} \int_{-\infty}^{\infty} \frac{\sqrt{n_j}}{\sqrt{2\pi\sigma^2}} e^{-\frac{n_j}{2\sigma^2}(\bar{x}_j-\theta)^2} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}(\theta-\mu)^2} d\theta.$$

Интеграл есть свёртка двух нормальных плотностей. Поэтому этот интеграл равен плотности нормального $\mathcal{N}(\mu, \tau^2 + \sigma^2/n_j)$ распределения. Следовательно,

$$L(\mu, \tau^2, \sigma^2 | x_j^{(n_j)}) = \frac{1}{(\sqrt{2\pi\sigma^2})^{n_j-1}} \frac{(2\pi)^{-1/2}}{\sqrt{\sigma^2 + n_j\tau^2}} \exp\left(-\frac{SS_j}{2\sigma^2} - \frac{n_j(\bar{x}_j - \mu)^2}{2(\sigma^2 + n_j\tau^2)}\right).$$

Таким образом, для всей матрицы данных $\mathbf{x}^{(k)}$

$$\begin{aligned} L(\mu, \tau^2, \sigma^2 | \mathbf{x}^{(k)}) &= \\ &= \frac{1}{(\sqrt{2\pi})^N \sigma^{N-k}} \exp\left(-\frac{SS}{2\sigma^2} - \frac{1}{2} \sum_1^k \frac{n_j(\bar{x}_j - \mu)^2}{\sigma^2 + n_j\tau^2} - \frac{1}{2} \sum_1^k \ln(\sigma^2 + n_j\tau^2)\right), \end{aligned}$$

где $N = \sum_1^k n_j$, $SS = \sum_1^k SS_j$. Следовательно, достаточная статистика включает сумму квадратов отклонений $SS = \sum_j \sum_i (x_{ji} - \bar{x}_j)^2$, а также средние арифметические \bar{x}_j во всех экспериментах. Понятно, что для полного восстановления функции правдоподобия необходимо также сохранять значения объёмов выборок в каждом эксперименте n_1, \dots, n_k .

Если объёмы выборок во всех экспериментах одинаковы: $n_j = n$, $j =$

$= 1, \dots, k$, то правдоподобие

$$L(\mu, \tau^2, \sigma^2 | \mathbf{x}^{(k)}) = \frac{1}{(\sqrt{2\pi})^N \sigma^{N-k}} \exp \left(-\frac{SS}{2\sigma^2} - \frac{n}{2(\sigma^2 + n\tau^2)} \sum_1^k (\bar{x}_j - \mu)^2 - \frac{k}{2} \ln(\sigma^2 + n\tau^2) \right),$$

Так как $\sum_1^k (\bar{x}_j - \mu)^2 = k(\bar{X} - \mu)^2 + SS_{\bar{x}}$, где $\bar{X} = \frac{1}{N} \sum_{ji} x_{ji}$, $SS_{\bar{x}} = \sum_1^k (\bar{x}_j - \bar{X})^2$, то достаточная статистика состоит из трёх компонент $(SS, SS_{\bar{x}}, \bar{X})$.

Ясно, что в этом случае оценка максимального правдоподобия для μ равна \bar{X} . Кроме того, легко видеть, что оценки максимального правдоподобия для σ^2 и τ^2 равны, соответственно,

$$\tilde{\sigma}^2 = \frac{1}{k(n-1)} \sum_{j=1}^k \sum_{i=1}^n (x_{ji} - \bar{x}_j)^2, \quad \tilde{\tau}^2 = \frac{1}{k} \sum_1^k (\bar{x}_j - \bar{X})^2 - \frac{1}{n} \tilde{\sigma}^2, \quad (1.67)$$

если $\tilde{\tau}^2 > 0$. В случае $\tilde{\tau}^2 \leq 0$ максимум функции правдоподобия достигается при $\tau^2 = 0$ в точке

$$\tilde{\sigma}_0^2 = \frac{1}{kn} \sum_{j=1}^k \sum_{i=1}^n (x_{ji} - \bar{x}_j)^2 + \frac{1}{k} \sum_1^k (\bar{x}_j - \bar{X})^2.$$

В классических монографиях по дисперсионному анализу (см., например, [25]) обычно рассматривают только оценки (1.67), которые совпадают с оценками по методу моментов.

В рамках теории дисперсионного анализа оценка $\tilde{\sigma}^2$ есть так называемая средняя сумма квадратов ошибок \overline{SS}_e , статистика $\overline{SS}_{\bar{x}} = \frac{1}{k-1} \sum_1^k (\bar{x}_j - \bar{X})^2$ — средняя сумма квадратов ошибок, обусловленная изменчивостью параметра θ . Отношение $\sigma^2 \overline{SS}_{\bar{x}} / ((\sigma^2 + n\tau^2) \overline{SS}_e)$ имеет распределение Фишера $F_{(k-1, k(n-1))}$ с $(k-1, k(n-1))$ степенями свободы (см. [25, формула (7.2.10)]). Таким образом, вероятность получения отрицательного значения оценки $\tilde{\tau}^2$ равна

$$\begin{aligned} \mathbf{P}(\tilde{\tau}^2 \leq 0) &= \mathbf{P} \left(\frac{\overline{SS}_A \sigma^2}{(\sigma^2 + n\tau^2) \overline{SS}_e} \leq \frac{k\sigma^2}{(\sigma^2 + n\tau^2)(k-1)} \right) = \\ &= F_{(k-1, k(n-1))} \left(\frac{k\sigma^2}{(\sigma^2 + n\tau^2)(k-1)} \right). \end{aligned}$$

При $k \rightarrow \infty$ вероятность $\mathbf{P}(\tilde{\tau}^2 \rightarrow 0)$, поэтому можно, воспользовавшись теоремой об асимптотической нормальности оценок максимального правдоподобия, построить доверительные интервалы для параметров модели. Для этого

необходимо уметь вычислять дисперсии оценок максимального правдоподобия. Как известно, эти дисперсии можно вычислить, если найти диагональные элементы обратной информационной матрицы Фишера. Из представления для функции правдоподобия легко найти, что информационная матрица Фишера

$$I(\mu, \sigma^2, \tau^2) = k \begin{pmatrix} \frac{n}{\sigma^2 + n\tau^2} & 0 & 0 \\ 0 & \frac{n-1}{2\sigma^4} + \frac{1}{2(\sigma^2 + n\tau^2)^2} & \frac{n}{2(\sigma^2 + n\tau^2)^2} \\ 0 & \frac{n}{2(\sigma^2 + n\tau^2)^2} & \frac{n^2}{2(\sigma^2 + n\tau^2)^2} \end{pmatrix}.$$

Обратная информационная матрица также легко вычисляется:

$$I^{-1}(\mu, \sigma^2, \tau^2) = \frac{1}{k} \begin{pmatrix} \tau^2 + \frac{\sigma^2}{n} & 0 & 0 \\ 0 & \frac{2\sigma^4}{n-1} & -\frac{2\sigma^4}{n(n-1)} \\ 0 & -\frac{2\sigma^4}{n(n-1)} & 2\tau^4 + \frac{2\sigma^2(\sigma^2 + 2(n-1)\tau^2)}{n(n-1)} \end{pmatrix}. \quad (1.68)$$

Подставляя сюда оценки, полученные в эксперименте, и построив на основе этих оценок доверительные интервалы надёжности $(1 - \alpha/3)$ для каждого из параметров, получим $(1 - \alpha)$ -доверительную область (параллелепипед) для всего трёхмерного вектора параметров. В свою очередь, это позволит построить соответствующие доверительные области для величин d-рисков статистических процедур.

Если объёмы выборок в архивных экспериментах не одинаковы, то для получения оценки максимального правдоподобия придётся воспользоваться каким-либо компьютерным средством нахождения максимума функции многих переменных.

Оценки параметров модели G-E. Предположим, что в эксперименте наблюдается выборка $X^{(n)}$ фиксированного объёма n из показательного распределения $\mathcal{E}(\theta)$ с плотностью $f(x|\theta) = \theta \exp\{-\theta x\}$, $x > 0$. Параметр θ есть реализация случайной величины ϑ с априорным гамма-распределением

$\mathcal{G}(\lambda, a)$ с плотностью

$$g(\theta; \lambda, a) = \frac{a^\lambda}{\Gamma(\lambda)} \theta^{\lambda-1} e^{-a\theta}, \quad \theta > 0.$$

Как следует из предыдущих рассуждений в архиве достаточно сохранять информацию только о значениях статистики $S_n = \sum_1^n X_i$ из каждого эксперимента. Безусловное распределение этой статистики приведено в лемме 1.4, секция 1.2. Обозначим распределение такого типа как $\mathcal{GE}(\lambda, a; n)$. Покажем, что это распределение определяется своими параметрами.

Теорема 1.8. *Если для какого-то $n \geq 1$ распределения $\mathcal{GE}(\lambda_1, a_1; n)$ и $\mathcal{GE}(\lambda_2, a_2; n)$ совпадают, то $\lambda_1 = \lambda_2$, $a_1 = a_2$.*

Доказательство. Пусть $\xi \sim \mathcal{GE}(\lambda, a; n)$. Условное (при заданном θ) распределение ξ совпадает с гамма-распределением с параметром формы n и параметром масштаба $1/\theta$. Легко видеть, что для $n \geq 2$ математическое ожидание $\mathbf{E}_\theta[\xi^{-1}] = \theta/(n-1)$. Следовательно, безусловное математическое ожидание

$$\mathbf{E}(\xi^{-1}) = \mathbf{E}(\mathbf{E}_\vartheta[\xi^{-1}]) = \frac{1}{n-1} \mathbf{E}\vartheta = \frac{\lambda}{a(n-1)}. \quad (1.69)$$

Плотность безусловного распределения ξ пропорциональна $t^{n-1}/(t+a)^{n+\lambda}$ (см. лемму 1.4). Ясно, что максимум этой плотности достигается в точке $t = (n-1)a/(\lambda+1)$. Следовательно, если $n \geq 2$ и распределения $\mathcal{GE}(\lambda_j, a_j; n)$, $j = 1, 2$, совпадают, то $\lambda_1/\alpha_1 = \lambda_2/\alpha_2$ и $a_1/(\lambda_1+1) = a_2/(\lambda_2+1)$, что, очевидно, возможно только при $a_1 = a_2$, $\lambda_1 = \lambda_2$.

Для $n = 1$ плотность распределения $\mathcal{GE}(\lambda, a; n)$ равна $\lambda a^\lambda/(t+a)^{\lambda+1}$. В точке $t = 0$ эта плотность равна λ/a . Кроме того, функция распределения $\mathcal{GE}(\lambda, a; n)$ в точке $t = a/\lambda$ равна $g(\lambda) = (1 + \lambda^{-1})^{-1}$, $\lambda > 0$. Очевидно, что функция $g(\lambda)$ есть возрастающая функция. Следовательно, совпадение распределений $\mathcal{GE}(\lambda_1, a_1; 1)$ и $\mathcal{GE}(\lambda_2, a_2; 1)$ возможно также только при $a_1 = a_2$, $\lambda_1 = \lambda_2$. \square

Момент $\mathbf{E}S_n^2$ существует только при $\lambda > 2$. Поэтому в общем случае нельзя построить оценки по методу моментов. Однако, аналогично (1.69) можно показать, что при $n > 2$ существует математическое ожидание

$$\mathbf{E} \frac{1}{S_n^2} = \frac{\lambda(\lambda+1)}{a^2(n-1)(n-2)}.$$

Таким образом, при $n > 2$ всегда можно построить оценки λ и a по методу «обратных моментов». Справедлива

Лемма 1.16. Пусть $M_{-j} = \sum_{i=1}^k t_i^{-j}/k$, $j = 1, 2$, — обратные моменты первого и второго порядков, вычисленные по k независимым реализациям t_1, \dots, t_k случайных величин $S_n \sim \mathcal{GE}(\lambda, a; n)$, $i = \overline{1, k}$, с некоторым фиксированным $n > 2$. Тогда оценки λ , a по методу обратных моментов вычисляются по формулам

$$\tilde{\lambda} = \frac{(n-1)M_{-1}^2}{(n-2)M_{-2} - (n-1)M_{-1}^2}, \quad \tilde{a} = \frac{M_{-1}}{(n-2)M_{-2} - (n-1)M_{-1}^2}.$$

Вероятностная модель $\mathcal{GE}(\lambda, a; n)$ имеет информационную матрицу Фишера следующего вида

$$I(\lambda, a; n) = \begin{pmatrix} \sum_{j=1}^n \frac{1}{(\lambda + j - 1)^2} & -\frac{n}{a(\lambda + n)} \\ -\frac{n}{a(\lambda + n)} & \frac{n\lambda}{a^2(\lambda + n + 1)} \end{pmatrix}. \quad (1.70)$$

Заметим, что, как и в N–N модели, здесь при увеличении $n \rightarrow \infty$ информационное количество не становится бесконечно большим, т.е. нельзя добиться сколь угодно высокой точности в оценках параметров за счёт увеличения только объёма испытаний n в каждом эксперименте — необходимо увеличение всего архива.

Для примера оценки по методу обратных моментов были смоделированы 10 000 раз при $\lambda = 3$, $a = 2$, $n = 10$, $k = 100$. Средние значения оценок оказались равными $\bar{\lambda} = 3.375$, $\bar{a} = 2.271$, т.е. наблюдается ненулевое смещение. Среднеквадратические ошибки оценок 1.103 и 0.589 соответственно. Одновременно, на одних и тех же данных, строились оценки максимального правдоподобия. Для этих оценок аналогичные характеристики оказались равными $\bar{\lambda} = 3.130$, $\bar{a} = 2.100$ со среднеквадратическими ошибками 0.335 и 0.198. Диагональные элементы обратной информационной матрицы (нормированные на k) равны (0.257, 0.151).

Если увеличить объём испытаний в каждом эксперименте до $n = 1000$, то получим $\bar{\lambda} = 3.118$, $\bar{a} = 2.084$ при среднеквадратических ошибках 0.262 и

0.135. Для оценок максимального правдоподобия снова получены более точные результаты: $\bar{\lambda} = 3.080$, $\bar{a} = 2.061$ со среднеквадратическими ошибками 0.185 и 0.098 соответственно. Диагональные элементы обратной информационной матрицы (нормированные на k) равны (0.163, 0.086).

Как и ожидалось, оценки максимального правдоподобия дают лучшие результаты.

Оценки параметров модели В–В. Предположим, что в эксперименте наблюдается случайная величина X , принимающая значения 1 и 0 с вероятностями θ и $1 - \theta$, соответственно, и в качестве априорного распределения выбрано бета-распределение с функцией плотности

$$g(\theta; p, q) = \frac{1}{B(p, q)} \theta^{p-1} (1 - \theta)^{q-1}, \quad 0 < \theta < 1,$$

где $B(p, q) = \Gamma(p)\Gamma(q)/\Gamma(p + q)$ — бета-функция Эйлера. Данная модель содержит два параметра (p, q) , неизвестные значения которых можно оценить по накопленному архиву данных.

Во-первых, заметим, что т.к. для модели Бернулли статистика $S_n = \sum X_k$ является достаточной, то архив будет представлять собой последовательность $\mathbf{s}^{(k)} = ((n_1, s_1), \dots, (n_k, s_k))$ значений статистики S_n в каждой выборке соответствующего объёма. Будем предполагать, что моменты остановок в каждом из архивных экспериментов не зависели от наблюдений и равнялись некоторому фиксированному числу n .

Покажем, что параметры В–В модели идентифицируемы по безусловному распределению достаточной статистики S_n . Будем говорить, что случайная величина ξ имеет бета-биномиальное распределение с параметрами (n, p, q) (кратко, $\xi \sim \mathcal{BB}(n, p, q)$), если ее функция распределения имеет вид (1.24), т.е. $\mathbf{P}(S_n = k) = C_n^k B(p + k, q + n - k)/B(p, q)$, $k = 0, 1, \dots, n$. Напомним, что математическое ожидание любой измеримой функции от такой случайной величины можно вычислять последовательно — сначала относительно биномиального распределения с параметрами (n, θ) , затем относительно бета-распределения параметра θ .

Лемма 1.17. Пусть $\xi \sim \mathcal{BB}(n, p, q)$, $n \geq 1$, тогда математическое

ожидаение и дисперсия ξ равны, соответственно,

$$\begin{aligned}\mu &= \mathbf{E}\xi = \frac{np}{p+q}, \\ \sigma^2 &= \mathbf{E}[\xi^2] - \mu^2 = \frac{npq(n+p+q)}{(p+q)^2(1+p+q)}.\end{aligned}$$

Обратно, если $\xi \sim \mathcal{BB}(n, p, q)$, $n > 1$, и $\mathbf{E}\xi = \mu$, $\mathbf{E}[\xi^2] - \mu^2 = \sigma^2$, то

$$p = \mu \frac{\sigma^2 - \mu(n - \mu)}{\mu(n - \mu) - n\sigma^2}, \quad q = (n - \mu) \frac{\sigma^2 - \mu(n - \mu)}{\mu(n - \mu) - n\sigma^2}. \quad (1.71)$$

Доказательство. Как отмечено выше,

$$\mathbf{E}[\xi] = \mathbf{E}[\mathbf{E}_{\vartheta}\xi] = \mathbf{E}[n\vartheta] = \frac{np}{p+q}.$$

Аналогично, по формуле для моментов бета-закона

$$\mathbf{E}[\xi^2] = \mathbf{E}[\mathbf{E}_{\vartheta}\xi^2] = \mathbf{E}[n\vartheta - n\vartheta^2 + n^2\vartheta^2] = \frac{np}{p+q} + \frac{n(n-1)p(1+p)}{(p+q)(1+p+q)}.$$

Отсюда уже легко получается заявленное выражение для дисперсии.

Обратно, т.к. $nq/(p+q) = n - \mu$, то из формулы для дисперсии получаем тождество

$$\frac{\mu(n - \mu)(n + p + q)}{n(1 + p + q)} = \sigma^2,$$

откуда легко находим для $n > 1$

$$p + q = n \frac{\sigma^2 - \mu(n - \mu)}{\mu(n - \mu) - n\sigma^2}.$$

Так как $p = \mu(p+q)/n$, $q = (n - \mu)(p+q)/n$, то отсюда следует утверждение леммы. \square

Теорема 1.9. Пусть $\xi_j \sim \mathcal{BB}(n, p_j, q_j)$, $n > 1$, $j = 1, 2$. Если распределения случайных величин ξ_1 , ξ_2 совпадают, то $p_1 = p_2$, $q_1 = q_2$.

Доказательство. Утверждение теоремы очевидным образом следует из предыдущей леммы. \square

Рассмотрим ситуацию, когда в каждом из архивных экспериментов проводилось одинаковое число контрольных испытаний: $n_i = n$, $\forall i = \overline{1, k}$. Прежде вычислим информационную матрицу Фишера относительно параметров p, q . Логарифм функции правдоподобия для этой модели равен

$$\mathcal{L}(p, q | S_n = k) = \ln C_n^k + \ln \mathbf{B}(p+k, q+n-k) - \ln \mathbf{B}(p, q).$$

Таблица 1.6: Нижние границы для стандартных отклонений оценок параметров В–В модели

		$p = 0.1,$ $q = 10$	$p = 1,$ $q = 1$	$p = 1,$ $q = 5$	$p = 5,$ $q = 5$	$p = 5,$ $q = 10$	$p = 10,$ $q = 10$
$\sqrt{\mathbf{D}(\tilde{p})}$	$n = 2$	3.5	4.3	9.4	60.4	86.4	220.4
$\sqrt{\mathbf{D}(\tilde{q})}$		360.2	4.3	47.6	60.4	173.0	220.4
$\sqrt{\mathbf{D}(\tilde{p})}$	$n = 10$	0.77	1.8	2.6	14.9	18.7	44.7
$\sqrt{\mathbf{D}(\tilde{q})}$		86.7	1.8	13.9	14.9	37.7	44.7
$\sqrt{\mathbf{D}(\tilde{p})}$	$n = 100$	0.28	1.4	1.4	7.7	8.0	16.1
$\sqrt{\mathbf{D}(\tilde{q})}$		41.6	1.4	8.3	7.7	16.4	16.1

Матрица вторых производных вычисляется через производную $\Psi'(x)$ дигамма-функции Эйлера $\Psi(x) = d \ln(\Gamma(x))/dx$. Легко видеть, что информационная матрица Фишера равна

$$I(p, q) = \begin{pmatrix} \tilde{\Psi}'(n, p, q) & \tilde{\Psi}'(n, p, q) \\ \tilde{\Psi}'(n, p, q) & \tilde{\Psi}'(n, p, q) \end{pmatrix} + \begin{pmatrix} \Psi'(p) - \mathbf{E}\Psi'(p + \xi) & 0 \\ 0 & \Psi'(q) - \mathbf{E}\Psi'(q + n - \xi) \end{pmatrix}, \quad (1.72)$$

где $\tilde{\Psi}'(n, p, q) = \Psi'(n + p + q) - \Psi'(p + q)$. В таблице 1.6 приведены значения диагональных элементов обратной информационной матрицы при некоторых значениях параметров. Как известно, диагональные элементы обратной информационной матрицы Фишера определяют нижнюю границу среднеквадратической ошибки оценок. Судя по приведённым значениям, для построения оценок с приемлемой точностью необходимо обладать архивом достаточно большого объёма ($k > 500$).

Заметим, что в рассматриваемой В–В модели число n входит как параметр модели и с увеличением этого параметра до $+\infty$ граница дисперсии оценок не будет уменьшаться до нуля (сравните с N–N моделью и G–E моделью).

Обсудим некоторые способы оценивания параметров В–В модели.

Опираясь на формулы (1.71) легко построить оценки по методу моментов. Проблема применения этих оценок заключается в том, что очень часто, особенно при малом значении объёма испытаний n или малом архиве k , эти оценки дают отрицательные значения, что неприемлемо.

Оценки максимального правдоподобия страдают похожим недугом. На численных реализациях метода максимального правдоподобия замечено, что в тех случаях, когда оценки по методу моментов дают отрицательные значения для p и q , максимум функции правдоподобия достигается при слишком больших значениях этих параметров (более 10^6).

После проведения большого числа модельных испытаний мы остановились на следующей модификации метода максимального правдоподобия.

Сначала находятся оценки (\tilde{p}, \tilde{q}) по методу моментов (1.71). Далее, начиная с $p_1 = \tilde{p}$, $q_1 = \tilde{q}$, проводится небольшое число $J = 4$ итераций поиска максимума логарифма функции правдоподобия по методу Ньютона. Очередное приближение по методу Ньютона вычисляется по формуле

$$(p_{j+1}, q_{j+1}) = (p_j, q_j) - \varkappa \mathcal{L}_2^{-1}(p_j, q_j | \mathbf{s}^{(k)}) \mathcal{L}'(p_j, q_j | \mathbf{s}^{(k)}), \quad j = 1, \dots, J,$$

где вектор первых производных вычисляется как

$$\begin{aligned} \mathcal{L}'(p, q | \mathbf{s}^{(k)}) &= \\ &= \frac{1}{k} \sum_{i=1}^k \begin{pmatrix} \Psi(p + s_i) \\ \Psi(q + n - s_i) \end{pmatrix} + \begin{pmatrix} \Psi(p + q) - \Psi(p) - \Psi(n + p + q) \\ \Psi(p + q) - \Psi(q) - \Psi(n + p + q) \end{pmatrix}. \end{aligned}$$

Матрица вторых моментов вычисляется аналогично информационной матрице Фишера (1.72) с заменой усреднения по случайной величине ξ на усреднение по архиву:

$$\begin{aligned} \mathcal{L}_2^{-1}(p_j, q_j | \mathbf{s}^{(k)}) &= \begin{pmatrix} \tilde{\Psi}'(n, p, q) & \tilde{\Psi}'(n, p, q) \\ \tilde{\Psi}'(n, p, q) & \tilde{\Psi}'(n, p, q) \end{pmatrix} + \\ &+ \frac{1}{k} \sum_{i=1}^k \begin{pmatrix} \Psi'(p) - \mathbf{E}\Psi'(p + s_i) & 0 \\ 0 & \Psi'(q) - \mathbf{E}\Psi'(q + n - s_i) \end{pmatrix}, \end{aligned}$$

где $\tilde{\Psi}'(n, p, q) = \Psi'(n + p + q) - \Psi'(p + q)$. Вместо суммирования компонент логарифмической функции правдоподобия мы взяли их усреднение по архиву, что не меняет сути метода. Если оценки по методу моментов отрицательны, то предлагается в качестве начальной точки итераций взять $p_1 = 1/\sqrt{k}$, $q_1 = lp_1$, где l — оценка отношения q/p :

$$l = \frac{n - \tilde{\mu}}{\tilde{\mu}}, \quad \tilde{\mu} = \frac{1}{k} \sum_{i=1}^k s_i.$$

Параметр \varkappa в очередной итерации подбирается так, чтобы оценки оставались положительными.

Для примера, было произведено 10^5 модельных испытаний предложенного способа оценивания при истинных значениях параметров $p = 0.1$, $q = 10$ и $n = 100$, $k = 500$. Средние значения оценок оказались равными $\bar{p} = 0.114$, $\bar{q} = 11.8$, медианы этих оценок $med(p) = 0.103$, $med(q) = 10.40$. Выборочные дисперсии оценок $s_p^2 = 0.00333$, $s_q^2 = 78.3$.

1.4.2 Непараметрические оценки плотности априорного распределения

В ситуации, когда нет оснований для выдвижения предположения о типе распределения неизвестного параметра, можно построить ядерную оценку этой плотности в рамках проблемы деконволюции. Теория таких оценок хорошо разработана (см., например, [31], [43]). Предполагается, что наблюдаемая в эксперименте случайная величина ξ представляет собой сумму независимых случайных величин: $\xi = \vartheta + \varepsilon$, причём плотность f распределения ε известна (может быть, с точностью до некоторого мешающего параметра ζ), а плотность g распределения ϑ требуется оценить по ряду независимых наблюдений $\mathbf{x}^{(k)} = (x_1, \dots, x_k)$, $k > 1$. Таким образом, плотность распределения ξ равна $f_\xi(x) = \int_{\mathbb{R}^1} f(x - \theta)g(\theta) d\theta$, функция распределения $F_\xi(x) = \int_{\mathbb{R}^1} F(x - \theta)g(\theta) d\theta$, где F — функция распределения ε .

В ситуации, когда мешающий параметр известен (или его можно достаточно точно оценить), в работе [31] предложена ядерная оценка с наилучшей скоростью сходимости следующего вида:

$$\hat{g}_k(\theta; \mathbf{x}^{(k)}) = \frac{1}{2\pi} \int_{-1/h}^{1/h} \varphi^{-1}(t)K(th)\psi_k(\theta, t; \mathbf{x}^{(k)}) dt,$$

где $\psi_k(\theta, t; \mathbf{x}^{(k)}) = \frac{1}{k} \sum_1^k \exp\{i(x_j - \theta)t\}$ — эмпирическая характеристическая функция, φ — характеристическая функция распределения с плотностью f . Ядро K можно выбирать произвольно, лишь бы функция $K(t)$ была чётной и достаточно быстро приближалась к 1 в точке $t = 0$. Наиболее часто рассматриваемые варианты $K(t) \equiv 1$ (равномерное ядро) или $K(t) = e^{-t^2/2}$ (гауссовское ядро). Ширину окна h в [31] предлагается выбирать, равной

$$h = h_k = \sqrt{2/\ln k}.$$

Функцию априорной плотности мы используем для вычисления рисков статистических процедур, представляющих собой интегральные характеристики. В связи с этим, предлагается выбор ширины окна связать с эмпирической функцией распределения $F_k(x; \mathbf{x}^{(k)})$ наблюдений x_1, \dots, x_k . Другими словами, величина окна h подбирается из условия минимизации расхождения между $F_k(x; \mathbf{x}^{(k)})$ и функцией распределения

$$\widehat{F}(x; \mathbf{x}^{(k)}) = \int_{\mathbb{R}^1} F(x - \theta) \widehat{g}_k(\theta; \mathbf{x}^{(k)}) d\theta. \quad (1.73)$$

Другие варианты выбора h (в том числе так называемый метод cross-validation) можно найти в [43], [36].

В статье [31] показано, что скорость, с которой проинтегрированная средняя квадратическая ошибка аппроксимации g с помощью \widehat{g}_k убывает к нулю, зависит от распределения f . В частности, для нормальной плотности f с известной дисперсией эта ошибка имеет порядок $O((\ln k)^{-r/2})$, где r — показатель гладкости «хвоста» характеристической функции априорного распределения (показатель пикообразности априорной плотности). Случай, когда ε имеет нормальное распределение с нулевым средним и стандартным отклонением ζ , которое оценивается также при помощи \widehat{g}_k , рассмотрен в [43].

Аналогичный подход используется и в ситуации, когда наблюдения в эксперименте представимы в виде произведения $\xi = \vartheta\varepsilon$, $\vartheta > 0$, $\varepsilon > 0$. В этом случае задача оценивания плотности ϑ сводится к задаче деконволюции с помощью перехода к логарифмам рассматриваемых величин: $\ln \xi = \ln \vartheta + \ln \varepsilon$.

Возможность применения ядерной оценки плотности для построения гарантийных процедур в d-апостериорном подходе устанавливается в следующей теореме. Будем считать мешающий параметр ζ известным и, кроме того, предположим, что функция распределения F непрерывна и строго возрастает. Обозначим через $R_0(c; q)$ d-риск второго рода, вычисленный относительно какой-либо априорной плотности q :

$$R_0(c; q) = \frac{1}{F_\xi(c)} \int_{\theta_0}^{\infty} F(c - \theta)q(\theta) d\theta,$$

где $F_\xi(c) = \int_{\mathbb{R}^1} F(c - \theta)q(\theta) d\theta$ — маргинальная функция распределения ξ ;

риск $R_0(c; q) = 0$, если $F_\xi(c) = 0$. Пусть $C(g)$ — единственное решение уравнения $R_0(c; g) = \beta_0$, а $C_k = \sup\{c \in \mathbb{R}^1 : R_0(c; \hat{g}_k) \leq \beta_0\}$.

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

$$\sup_{c \in Q_0} |R_0(c; \hat{g}_k) - R_0(c; g)| = O_{\mathbf{P}}(1/v_k), \text{ где } Q_0 = \{c : F_\xi(c) > 0\};$$

Доказательство. Напомним (см., например, [61]), что соотношение $\xi_k = O_{\mathbf{P}}(1/v_k)$, $k \rightarrow \infty$, для какой-либо последовательности случайных величин означает, что последовательность $v_k \xi_k$ ограничена по вероятности, то есть $\sup_{k \geq 1} \mathbf{P}(|v_k \xi_k| > x) \rightarrow 0$ при $x \rightarrow \infty$. Как показано в [21], в силу непрерывности функции распределения F риск $R_0(c; g)$ есть непрерывная монотонно возрастающая функция c . Поэтому для доказательства первой части теоремы достаточно доказать соответствующее утверждение для каждой точки c такой, что $F_\xi(c) > 0$.

Числитель и знаменатель функции риска (с априорной плотностью $q = g$ или $q = g_k$) представляют собой интегральные выражения вида

$$A_m(c; q) = \int_{-\infty}^{\infty} \chi_m(\theta) F(c - \theta) q(\theta) d\theta, \quad m = 0, 1,$$

где в знаменателе d-риска ($m = 1$) функция $\chi_1(\theta) \equiv 1$, а в числителе $\chi_0(\theta)$ есть индикаторная функция множества $[\theta_0, \infty)$. Поэтому для $m = 0, 1$

$$v_k \sup_c |A_m(c; g) - A_m(c; \hat{g}_k)| \leq v_k \int_{-\infty}^{\infty} |\hat{g}_k(\theta; \mathbf{X}^{(k)}) - g(\theta)| d\theta = O_{\mathbf{P}}(1).$$

Нормированная разность рисков

$$\begin{aligned} & v_k (R_0(c; \hat{g}_k) - R_0(c; g)) = \\ & = \frac{1}{A_1(g_k)} (A_0(g_k) - A_0(g)) v_k + \frac{A_0(g)}{A_1(g) A_1(g_k)} (A_1(g_k) - A_1(g)) v_k = O_{\mathbf{P}}(1), \end{aligned}$$

т.к. по условию теоремы $A_1(g_k) \rightarrow A_1(g) \neq 0$ при $k \rightarrow \infty$ по вероятности относительно безусловного распределения \mathbf{P} . \square

Ядерная оценка априорной плотности с нормальными наблюдениями. В нормально-нормальной модели наблюдение есть результат сложения двух независимых случайных величин: $X = \vartheta + \varepsilon$, где «ошибка» ε имеет нормальный закон распределения с нулевым средним и некоторой (неизвестной) дисперсией σ^2 . Если не предполагать нормальность распределения ϑ , но сохранить структуру наблюдаемой случайной величины в виде свёртки, то вполне резонно попытаться построить ядерную оценку плотности ϑ в рамках проблемы деконволюции. Для рассматриваемого здесь случая, когда дисперсия наблюдений σ^2 неизвестна, в статье [43] предлагается оценка априорной плотности g с одновременной оценкой дисперсии. Однако в отличие от ситуации [43] в нашем случае в каждом эксперименте проводится более одного наблюдения и архив будет содержать, кроме значений $\bar{x}_1, \dots, \bar{x}_k$ статистики выборочного среднего \bar{X} , значения несмещённых оценок дисперсии s_1^2, \dots, s_k^2 . Среднее арифметическое $\hat{\sigma}_k^2$ этих оценок даёт состоятельную (при $k \rightarrow \infty$) оценку σ^2 с максимально возможной скоростью $O(1/k)$ убывания к нулю среднеквадратической ошибки. Поэтому при построении оценки априорной плотности можно считать, что дисперсия наблюдений известна. В этой ситуации в работе [31] предложена ядерная оценка с наилучшей скоростью сходимости следующего вида:

$$\hat{g}_k(\theta; \mathbf{x}^{(k)}) = \frac{1}{\pi} \int_0^{1/h} \exp\left(\frac{t^2 \hat{\sigma}_k^2}{2}\right) K(th) \psi_k(\theta, t; \mathbf{x}^{(k)}) dt,$$

где $\psi_k(\theta, t; \mathbf{x}^{(k)}) = \frac{1}{k} \sum_1^k \cos((\bar{x}_j - \theta)t)$ — оценка действительной части характеристической функции наблюдаемой в эксперименте статистики \bar{X} , $\hat{\sigma}_k^2$ — оценка дисперсии. Ядро K можно выбирать произвольным образом, лишь бы функция $K(\theta)$ достаточно быстро приближалась к 1 в точке $\theta = 0$. Ширину окна в [31] предлагается выбирать, равной $h = h_k = \sqrt{2/\ln k}$.

Ядерная оценка плотности может принимать отрицательные значения. В силу этого возможна ситуация, когда для функции д-риска, вычисленной относительно такой плотности, не будет выполняться свойство монотонности по значениям критической константы. Поэтому предлагается использовать исправленную оценку $\tilde{g}_k(\theta; \mathbf{x}^{(k)}) = B \max(0, \hat{g}_k(\theta; \mathbf{x}^{(k)}))$ с нормирующим коэффициентом B , при котором $\int_{\mathbb{R}^1} \tilde{g}_k(\theta; \mathbf{x}^{(k)}) d\theta = 1$.

Как отмечено выше, при выборе ширины окна h следует стремиться к

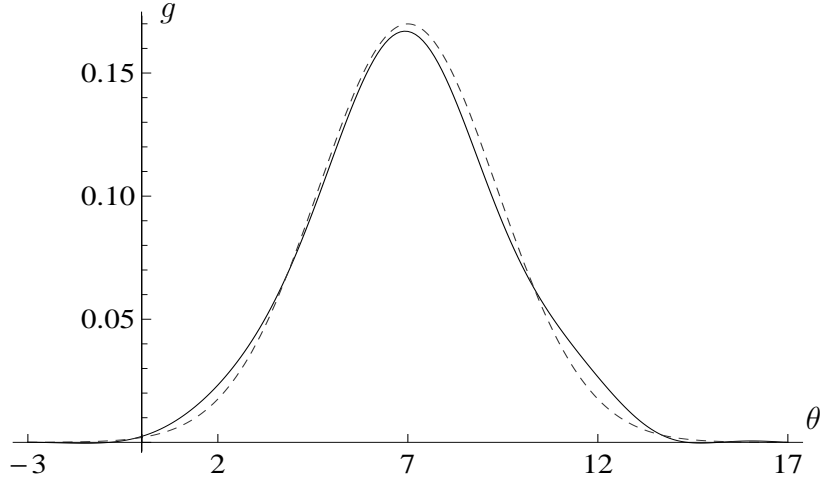


Рис. 1.4: Ядерная оценка априорной плотности

уменьшению расхождения между функцией (1.73) и эмпирической функцией распределения $F_k(x; n)$, построенной по значениям выборочных средних $\bar{x}_1, \dots, \bar{x}_k$. При нормальных наблюдениях величина окна h и область изменения $\theta \in (A, B)$ подбирается из условия минимизации расхождения между $F_k(x; n)$ и функцией распределения

$$\tilde{F}_k(x; n) = \int_A^B \Phi \left(\frac{x - \theta}{\hat{\sigma}_k} \sqrt{n} \right) \tilde{g}_k(\theta; \mathbf{x}^{(k)}) d\theta.$$

На рисунке 1.4 приведён график оценки априорной плотности (сплошная линия), построенной по $k = 700$ данным, описанным ниже в главе 2, в сравнении с параметрической оценкой в рамках N–N модели (пунктирная линия). Применялось равномерное ядро $K(t) \equiv 1$. Вместо рекомендованного авторами [31] окна $h = \sqrt{2/\ln k} \approx 0.553$ использовано $h = 0.575$, т.к. при таком окне уменьшается расхождение между эмпирическими функциями распределения статистик (см. пояснение в предыдущем абзаце). По тем же соображениям оценки априорной плотности находились для значений $\theta \in (A, B) = (-2.5, 16.5)$. Если для вычисления НОВ воспользоваться этой априорной плотностью (с оценкой $\tilde{\sigma}^2 = 1.95$), то для ограничений $\beta_0 = 0.1$, $\beta_1 = 0.001$ получим $n^* = 8$ (сравните с НОВ $n^* = 5$ для N–N модели); в случае более жёстких ограничений $\beta_0 = 0.05$, $\beta_1 = 0.0001$ различие в определении НОВ более существенное — $n^* = 311$ для ядерной оценки плотности против $n^* = 172$ для полностью параметрической N–N модели.

Глава 2

D-апостериорный подход к проблемам контроля качества и множественного тестирования генов

Применение математической статистики к задачам контроля качества и генетическим исследованиям во многом спровоцировали создание совершенно новых методов планирования статистических экспериментов и построения гарантийных процедур статистического вывода. Монография Вальда [5], посвященная методам последовательного планирования наблюдений с целью принятия одной из гипотез с заданными ограничениями на вероятностями ошибочных решений, изначально была тесно связана с проблемой гарантийности выборочных методов приемочного контроля. С другой стороны, по-видимому, именно этот подход к планированию испытаний привел Вальда к созданию аксиоматической теории построения процедур статистических решений (см. [5], стр. 308–325). Применение методов множественного тестирования гипотез к задачам генетики дало толчок другому, отличному от классического, подходу к проблеме гарантийности статистического вывода с введением таких характеристик процедуры тестирования как FWER, FDR, FNR и т.п. (см. [32]).

Наш подход к решению такого рода статистических проблем в первую очередь основан на использовании байесовской парадигмы. Изменчивость контролируемой характеристики при контроле каждой партии продукции носит, очевидно, случайный характер, так что существование априорного распределения не вызывает сомнения. Тестируемый параметр в задаче проверки большого числа гипотез по данным экспрессии генов также принимает различные значения для каждого гена, так что отрицать в данной ситуации наличие априорного распределения сродни утверждению о невозможности применения вероятностных методов в исследованиях биологических популяций. Как отмечалось во введении диссертации игнорирование априорного распределения в построении гарантийных процедур приводит не только к увеличению необходимого числа наблюдения, но и не позволяет контролировать именно те характеристики процедур, которые в наибольшей степени соответствуют существу решаемой

статистической проблемы. Предлагаемый нами d -апостериорный подход к проблеме гарантийности статистического вывода во многом снимает такого рода проблемы, когда решаются задачи гарантийного контроля качества и аттестации выпускаемой продукции, а также задачи множественного тестирования гипотез с целью выявления генов, ответственных за патологию организма, выявления медицинских показателей, в наибольшей степени влияющих на точность диагноза, и т.п.

В первой секции данной главы d -гарантийные статистические процедуры, разработанные в первой главе, применяются к задачам контроля качества и сравниваются по объёму выборок при различных параметрах вероятностной модели. Основная цель — показать преимущества последовательных схем и выяснить границы их применимости. Во второй секции методика построения d -гарантийных процедур применяется к проблеме множественного тестирования, в частности к проблеме выделения из большого числа генов тех из них, которые показывают изменение экспрессии в группе заболевших пациентов. Показывается, что в рамках d -апостериорного подхода можно контролировать оптимальным образом не только d -риск первого рода (байесовский аналог $pFDR$), но и d -риск второго рода (аналог $pFNR$), а также строить процедуры различения более двух гипотез.

2.1 Статистический контроль качества

Основная цель статистического контроля качества состоит в понижении доли некондиционной продукции, отсылаемой потребителю, до приемлемого уровня, оговоренного в обязательствах производителя продукции перед потребителем. Решение такого рода задач осуществляется с помощью контроля части производимой продукции — так называемого выборочного или статистического контроля качества. В этом случае с некоторой вероятностью α (называемой обычно вероятностью ошибки первого рода или риском изготовителя) процедура контроля может пропустить некондиционный продукт или, наоборот, с вероятностью β (называемой вероятностью ошибки второго рода или риском потребителя) отклонить кондиционный. Пусть Q_{in} — входной

уровень качества, то есть доля кондиционной продукции среди изготовленной предприятием; необходимость контроля продукции обуславливается как правило невысоким значением Q_{in} , которое не удовлетворяет потребителя. Зная значения Q_{in} , α и β , можно, используя формулу Байеса, вычислить выходной уровень качества Q_{out} — долю кондиционной продукции, среди отсылаемой потребителю после контроля:

$$Q_{out} = \frac{(1 - \alpha)Q_{in}}{(1 - \alpha)Q_{in} + \beta(1 - Q_{in})}.$$

Кроме выходного уровня качества существует ещё один аспект данной проблемы, который важен для изготовителя продукции. Его, естественно, интересует доля некондиционной продукции среди той, что была отклонена при контроле. Это так называемый уровень контроля Q_{cont} , который также легко вычисляется по формуле Байеса:

$$Q_{cont} = \frac{(1 - \beta)(1 - Q_{in})}{\alpha Q_{in} + (1 - \beta)(1 - Q_{in})}.$$

Такая простая связь между рисками потребителя и изготовителя с одной стороны и выходными характеристиками контроля с другой стороны возникает только в случае, когда контроль качества сводится к различению двух простых гипотез. В обычной практике применения статистических методов контроля вопрос о кондиционности партии продукции решается с помощью статистических критериев различения сложных гипотез. Для таких критериев чаще всего удаётся контролировать лишь одну из вероятностей ошибок — вторая из ошибок будет иметь малые значения только вне некоторой области «безразличия», причём обе эти ошибки существенным образом зависят от контролируемой характеристики качества. Но в таком случае, приведённые формулы для вычисления Q_{out} и Q_{cont} уже не работают.

Существуют методы так называемых последующих оценок выходного качества на основе накопленного архива данных предыдущих инспекций выпускаемой продукции (см. монографию Ю.К. Беляева [1]). После оценки Q_{out} и Q_{cont} рекомендуется изменить риск потребителя и/или изготовителя с тем, чтобы выходной уровень качества стал приемлемым. По-сути, эти методы ориентированы на контроль классических вероятностей ошибок первого и второго рода в ситуации, когда от процедуры контроля требуется удовлетворение

ограничений на выходной уровень качества Q_{out} и уровень контроля Q_{cont} . При этом негласно предполагается, что в последовательности статистических экспериментов контролируемая характеристика изменяется случайно. Вполне естественно поставить задачу оптимизации статистического вывода с гарантиями на уровне качества и/или контроля. Решить такую задачу можно в рамках d-апостериорного подхода к проблеме гарантийного вывода.

Наша цель сравнить изученные в первой главе процедуры d-гарантийного тестирования гипотез с целью выявления наиболее экономного с точки зрения среднего объёма наблюдений статистического критерия. Сравнения будут проводиться в рамках конкретных вероятностных моделей, наиболее часто используемых при контроле качества выпускаемой продукции. Например, при контроле плавки металла по предельной прочности или контроле дизельного топлива по общему содержанию серы (контроль за содержанием вредных примесей, в общем случае) вероятностная модель N–N имеет теоретическое обоснование в рамках центральной предельной теоремы и давно апробирована на реальных данных. При контроле по альтернативному признаку использование модели В–В вообще не вызывает сомнения. Вероятностная модель G–E обычно применяется при контроле надёжности радиотехнических изделий. Сравнение объёмов наблюдений будет проводиться, в основном, на данных статистического моделирования

2.1.1 Выбор процедуры контроля качества в рамках модели N–N

Пусть требуется проверить гипотезу $H_0 : \theta \leq \theta_0$ против альтернативы $H_1 : \theta > \theta_0$ в рамках нормально-нормальной модели (1.10). Как отмечено ранее, сравнение процедур проверки гипотез достаточно проводить при параметрах $\sigma^2 = \tau^2 = 1$ и $\theta_0 = 0$. Границы продолжения эксперимента для процедур первого перескока и на статистике вклада приведены в (1.60), (1.62). В таблицах 2.1, 2.2 рассматривается усечённый вариант процедуры первого перескока, когда эксперимент останавливается с вероятностью единица после проведения $n = n^*$ наблюдений, где n^* — необходимый объём выборки процедуры с фиксированным числом наблюдений (также приведён в таблицах). При этом решение в пользу гипотезы H_0 на шаге $n = n^*$ принимается, если

выборочное среднее $\bar{x} \leq (c_0 + c_1)/2$, где константы $c_0 \geq c_1$ найдены из уравнений $\mathcal{R}_j(n^*, c_j) = \beta_j$, $j = 0, 1$. Оценки характеристик процедуры строились по $k = 10^5 \div 10^6$ репликациям метода стохастического моделирования.

Из приведённых таблиц можно сделать следующие

Выводы 2.1. Для нормально-нормальной модели имеет место:

1. При умеренной величине превышения требуемого уровня надёжности усечённая процедура значительно понижает среднее количество проводимых экспериментов.

2. Момент остановки усечённой процедуры в редких случаях (от 1.5% до 4%) достигает крайнего значения n^* .

3. Последовательная процедура на статистике вклада понижает средний объём наблюдений чуть в меньшей степени, чем усечённая универсальная процедура.

4. d-риск последовательной процедуры на статистике вклада не сильно отличается от номинальных уровней β_0, β_1 только при равных ограничениях и при априорном среднем, близком к границе различаемых гипотез — чем меньше вероятность какой-либо гипотезы, тем больше превышение соответствующего d-риска над номинальным уровнем.

5. Оценки показателя степени правого хвоста распределения ν_{un} всегда меньше единицы, что подтверждает предположение о бесконечном значении математического ожидания $\mathbf{E}[\nu_{un}]$ момента остановки универсальной d-гарантийной процедуры.

Контроль стальных дисков по предельной прочности. Различие в свойствах рассмотренных трёх схем проведения эксперимента продемонстрируем на примере реальных данных, собранных на одном из промышленных предприятий г. Казани. При контроле качества выпускаемого изделия измеряется предельная прочность у нескольких образцов, снятых с припуска изделия. Предельная прочность образца определяется как наименьшая нагрузка (в кг. на мм^2), при которой происходит его разрыв. Цель контроля состоит в принятии испытуемого изделия к эксплуатации или отклонении его от приёмки. Предельная прочность изделия θ связана с предельной прочностью об-

Таблица 2.1: Характеристики последовательных процедур; N–N модель

	$\beta_1 = \beta_0 = 0.01$			$\beta_0 = 0.001, \beta_1 = 0.005$		
μ	2	1	0	2	1	0
Π_0	0.023	0.159	0.5	0.023	0.159	0.5
n^*	51	540	1 013	417	5 617	16 059
$(n_\beta - n^*, \tilde{n}_\beta - n^*)$	(15, 4)	(6, 1)	(1, 1)	(61, 5)	(66, 1)	(37, 1)
Процедура первого перескока, усечённая на уровне n^*						
\mathcal{R}_0	0.012	0.014	0.014	0.0015	0.0014	0.0016
\mathcal{R}_1	0.010	0.012	0.014	0.0056	0.007	0.008
$\mathbf{E} \tilde{\nu}_{un}$	3.5	38.2	84.3	10.1	159	453
$\mathbf{E}[\tilde{\nu}_{un} \vartheta \leq 0]$	32.6	108	83.3	169	515	505
$\mathbf{E}[\tilde{\nu}_{un} \vartheta > 0]$	2.84	25.6	85.3	6.49	91.5	401
медиана $\tilde{\nu}_{un}$	1	4	10	1	5	15
$\mathbf{P}(\nu_{un} \geq n^*)$	0.024	0.033	0.038	0.011	0.014	0.013
$\Psi(d_0)$	0.013	0.151	0.500	0.017	0.155	0.495
$\mathbf{P}(\nu_{un} \geq x)$	$\frac{0.22}{x^{0.57}}$	$\frac{1.12}{x^{0.56}}$	$\frac{0.60}{x^{0.56}}$	$\frac{0.27}{x^{0.55}}$	$\frac{1.21}{x^{0.52}}$	$\frac{2.10}{x^{0.51}}$
Процедура на статистике вклада						
\mathcal{R}_0	0.013	0.0095	0.0097	0.0027	0.0012	0.0012
\mathcal{R}_1	0.006	0.0095	0.0097	0.0038	0.0049	0.0049
$\mathbf{E} \nu_{sc}$	4.26	41.7	92.6	14.5	197	532
$\mathbf{E}[\nu_{sc} \vartheta \leq 0]$	62.9	134.8	92.2	395	906	720
$\mathbf{E}[\nu_{sc} \vartheta > 0]$	2.87	24.0	93.1	5.85	65.3	342
медиана ν_{sc}	2	13	40	3	28	172
99%-квантиль ν_{sc}	61	490	939	297	3076	6961
$\mathbf{P}(\nu_{un} \geq n^*)$	0.012	0.008	0.009	0.007	0.004	0.003
$\Psi(d_0)$	0.018	0.152	0.500	0.019	0.153	0.501

Таблица 2.2: Характеристики последовательных процедур; N–N модель

	$\beta_1 = \beta_0 = 0.05$			$\beta_0 = 0.01, \beta_1 = 0.05$		
μ	1.5	1	0	1.5	1	0
Π_0	0.067	0.159	0.5	0.067	0.159	0.5
n^*	6	21	40	11	49	157
$(n_\beta - n^*, \tilde{n}_b - n^*)$	(4, 2)	(1, 1)	(1, 1)	(10, 4)	(8, 2)	(4, 1)
Процедура первого перескока, усечённая на уровне n^*						
\mathcal{R}_0	0.051	0.054	0.058	0.010	0.012	0.012
\mathcal{R}_1	0.050	0.053	0.058	0.050	0.056	0.062
$\mathbf{E} \tilde{\nu}_{un}$	1.94	5.21	11.4	2.43	8.50	27.2
$\mathbf{E}[\tilde{\nu}_{un} \vartheta \leq 0]$	4.89	11.7	11.4	8.48	24.5	32.4
$\mathbf{E}[\tilde{\nu}_{un} \vartheta > 0]$	1.73	4.0	11.4	2.00	5.50	22.0
медиана $\tilde{\nu}_{un}$	1	2	5	1	2	7
$\mathbf{P}(\nu_{un} \geq n^*)$	0.118	0.107	0.125	0.085	0.080	0.077
$\Psi(d_0)$	0.019	0.151	0.500	0.017	0.112	0.474
Процедура на статистике вклада						
\mathcal{R}_0	0.101	0.069	0.049	0.022	0.015	0.011
\mathcal{R}_1	0.024	0.035	0.049	0.024	0.036	0.047
$\mathbf{E} \nu_{sc}$	2.53	5.60	11.8	3.88	10.6	30.7
$\mathbf{E}[\nu_{sc} \vartheta \leq 0]$	9.37	13.3	11.8	24.8	38.1	40.1
$\mathbf{E}[\nu_{sc} \vartheta > 0]$	2.04	4.15	11.8	2.37	5.49	21.3
медиана ν_{sc}	1	3	7	1	3	16
99%-квантиль ν_{sc}	21	41	69	52	117	236
$\mathbf{P}(\nu_{sc} \geq n^*)$	0.087	0.048	0.045	0.067	0.048	0.026
$\Psi(d_0)$	0.049	0.138	0.499	0.045	0.129	0.482

разца X соотношением $X = \theta + \varepsilon$, где ε — случайная «ошибка» в измерении прочности, обусловленная отклонениями от стандарта в изготовлении образца, локальными дефектами металла в теле припуска, неравномерностью подачи нагрузки, ошибкой показания прибора в момент разрыва образца. Согласно общему положению Теории Ошибок, распределение ε должно подчиняться нормальному закону с нулевым средним и некоторой дисперсией σ^2 . По результатам измерений образцов необходимо принять решение о кондиционности изделия, т.е. проверить гипотезу $H_1: \theta \geq \theta_0$ о значении прочности изделия, где θ_0 — норма ГОСТа.

Значение θ для каждого изделия есть также реализация некоторой случайной величины, распределение которой определяется варьированием общей прочности металла, из которого изготавливается изделие. Прочность легированной стали определяется, в основном, отсутствием дефектов в кристаллической структуре металла. Если предположить, что число кристаллов с нарушенной структурой его решетки подчиняется биномиальному закону, то предположение нормальности распределения ϑ становится вполне естественным.

Мы располагали данными испытаний более 700 изделий, изготовленных в течение одного года. У каждого изделия замерялась предельная прочность X_1, X_2 двух образцов ($n = 2$), взятых из припуска изделия. Проверка на нормальность случайных величин X_1, X_2 по выборке объёма $k = 700$ с помощью критерия хи-квадрат указывало на хорошее согласие с гипотезой нормальности. В силу теоремы Г. Крамера нормальность распределения суммы влечет нормальность распределения как X , так и ϑ .

Столь большой архив данных позволил достаточно точно оценить параметры μ, τ^2, σ^2 в рамках модели II дисперсионного анализа (см. секцию 1.4.1). Оценки максимального правдоподобия для σ^2, τ^2 оказались равными $\tilde{\sigma}^2 = 1.95, \tilde{\tau}^2 = 5.52$, соответственно. Оценка для априорного среднего $\tilde{\mu} = 7.0$ (после вычитания нормы θ_0). Диагональные элементы обратной информационной матрицы для этой модели (см. (1.68)), которые являются оценками для дисперсий оценок максимального правдоподобия, равны соответственно

$$D\tilde{\mu} = 0.0093, \quad D\tilde{\sigma}^2 = 0.0108, \quad D\tilde{\tau}^2 = 0.1232.$$

Построив на основе этих оценок для каждого из параметров доверительный

Таблица 2.3: Сравнение трёх схем выбора на данных реального примера

$\mu = 7$ $\sigma^2 = 1.95, \tau^2 = 5.52$ $\theta_0 = 0$ $\mathbf{P}(\vartheta < 0) = 0.00144$	$\beta_0 = 0.1, \beta_1 = 0.001$ $n^* = 5,$ $(n_\beta = 7, \tilde{n}_\beta = 6)$		$\beta_0 = 0.05, \beta_1 = 0.0001$ $n^* = 172,$ $(n_\beta = 172, \tilde{n}_\beta = 172)$	
	$\hat{\varphi}_{un}$	φ_{sc}	$\hat{\varphi}_{un}$	φ_{sc}
\mathcal{R}_0	0.12	0.27	0.072	0.078
\mathcal{R}_1	0.00098	0.0005	0.000084	0.00011
$\Psi(d_0)$	0.00056	0.0013	0.0015	0.0014
$\mathbf{E}[\nu]$	1.19	1.05	1.73	2.31
$\mathbf{E}[\nu \vartheta < \theta_0]$	4.41	4.10	46.4	32.8
$\mathbf{E}[\nu \vartheta \geq \theta_0]$	1.17	1.02	1.66	2.27
Медиана ν	1	1	1	2
0.99-квантиль ν	5	2	10	7

интервал надёжности $1 - 0.01/3$, получим приближённую 99 %-ю доверительную область (параллелепипед) для всего трёхмерного вектора параметров модели. Границы этого параллелепипеда выглядят так:

$$\mu \in (6.74, 7.26), \quad \sigma^2 \in (1.67, 2.23), \quad \tau^2 \in (4.57, 6.47).$$

При параметрах модели, совпадающих с указанными оценками, уровень входного качества $1 - \Pi_0 = 0.9986$ очень высок, что позволяет построить процедуру контроля с приемлемым объёмом инспекций и жёсткими ограничениями ($\beta_1 = 10^{-4}$) на долю некондиционных изделий среди отосланных потребителю, т.е. на величину условной вероятности $\mathbf{P}(\vartheta < \theta_0 | \delta = d_1)$.

Следует сказать, что на предприятии изделие считалось кондиционным при контроле, если наименьшее значение прочности образцов $\min(x_1, x_2) > 1$. Значение d-риска второго рода такой процедуры малò:

$$\mathbf{P}(\vartheta \leq 0 | \min(X_1, X_2) > 1) = \frac{1}{\tau \Psi(d_1)} \int_{-\infty}^0 \Phi^2\left(\frac{\theta - 1}{\sigma}\right) \phi\left(\frac{\theta - \mu}{\tau}\right) d\theta = 3 \cdot 10^{-5},$$

где $\Psi(d_1) = \int_{-\infty}^{\infty} \Phi^2((\theta - 1)/\sigma) \phi((\theta - \mu)/\tau) d\theta/\tau$. В то время как d-риск первого рода $\mathbf{P}(\vartheta > 0 | \min(X_1, X_2) \leq 1) = 0.94$. Расчёты показывают, что необходимо провести более 630 измерений образцов, чтобы процедура, основанная на минимальном значении выборки, имела d-риски $\mathcal{R}_0 = 0.1, \mathcal{R}_1 = 0.001$. Другими словами, эффективность ранговых тестов здесь крайне мала.

В таблице 2.3 приведены характеристики всех трёх предложенных здесь схем контроля при описанных выше параметрах модели. Число шагов универсальной последовательной процедуры (обозначено $\widehat{\varphi}_{un}$) ограничивалось НОВ n^* . Последовательная процедура на статистике вклада обозначена φ_{sc} . Данные таблицы указывают на высокую эффективность последовательных процедур контроля в сравнении с контролем по фиксированному объёму испытаний: средний объём наблюдений на порядок меньше НОВ. Следует особо отметить, что среднее значение объёма наблюдений $\mathbf{E}\{\nu \mid \theta > \theta_0\}$ в случае контроля кондиционной продукции значительно меньше среднего значения $\mathbf{E}\{\nu \mid \theta \leq \theta_0\}$ в случае контроля «плохой» продукции.

2.1.2 Приёмочный контроль в рамках модели G-E

Пусть требуется проверить гипотезу $H_0 : \theta \leq \theta_0$ против альтернативы $H_1 : \theta > \theta_0$ в рамках гамма-показательной модели (1.18). Снова сравнение процедур проверки гипотез достаточно проводить при $\theta_0 = 1$. Границы продолжения эксперимента для процедур первого перескока и на статистике вклада приведены в (1.63), (1.64). В таблице 2.4 рассматривается усечённый вариант процедуры первого перескока, когда эксперимент останавливается с вероятностью единица после проведения $n = n^*$ наблюдений, где n^* — необходимый объём выборки процедуры с фиксированным числом наблюдений (также приведён в таблицах). При этом — решение в пользу нулевой гипотезы при $\nu_{un} = n = n^*$ принимается, когда сумма выборочных данных S_n удовлетворяет неравенству $S_n / (S_n + a) > \mathbb{B}^{-1}(Q; n, \lambda)$, где $Q = (1 - \Pi_0 - \beta_0) / (1 - \beta_0 - \beta_1)$ (см. (1.20)).

Оценки характеристик процедуры строились по $k = 10^5 \div 10^6$ репликациям метода стохастического моделирования. Параметры (λ, a) выбирались так, чтобы априорная вероятность нулевой гипотезы $\Pi_0 \in (\beta_1, 1 - \beta_0)$. В противном случае одну из гипотез можно принимать всегда без проведения наблюдений.

Результаты модельных испытаний в рамках модели G-E позволяют сделать вывод, что свойства последовательных схем здесь аналогичны свойствам для модели N-N.

Таблица 2.4: Характеристики последовательных процедур; G–E модель

	$\beta_1 = \beta_0 = 0.10$				$\beta_0 = \beta_1 = 0.05$		
(λ, a)	(1, 0.5)	(1, 1)	(3, 4)	(35, 30)	(1, 0.5)	(1, 1)	(3, 4)
Π_0	0.393	0.632	0.762	0.203	0.393	0.632	0.762
n^*	6	10	26	166	24	37	107
$(n_\beta - n^*, \tilde{n}_b - n^*)$	(1, 0)	(0, 0)	(1, 2)	(44, 30)	(1, 0)	(0, 0)	(0, 2)
$n_\tau - n^*$	2	0	-3	20	10	3	-10
Процедура первого перескока, усечённая на уровне n^*							
\mathcal{R}_0	0.091	0.095	0.102	0.115	0.050	0.051	0.055
\mathcal{R}_1	0.108	0.112	0.116	0.124	0.062	0.062	0.067
$\mathbf{E} \tilde{\nu}_{un}$	3.37	4.74	9.54	42.7	7.60	10.8	25.3
$\mathbf{E}[\tilde{\nu}_{un} \vartheta \leq 1]$	3.20	3.72	7.09	76.6	7.68	8.20	17.9
$\mathbf{E}[\tilde{\nu}_{un} \vartheta > 1]$	3.48	6.52	17.34	34.0	7.56	15.3	49.2
медиана $\tilde{\nu}_{un}$	3	4	5	17	4	5	9
$\mathbf{P}(\nu_{un} \geq n^*)$	0.262	0.227	0.173	0.106	0.133	0.128	0.106
$\Psi(d_0)$	0.357	0.657	0.826	0.107	0.372	0.640	0.791
Процедура на статистике вклада							
\mathcal{R}_0	0.080	0.065	0.065	0.164	0.045	0.040	0.039
\mathcal{R}_1	0.094	0.127	0.162	0.077	0.049	0.059	0.067
$\mathbf{E} \nu_{sc}$	3.42	4.50	9.62	60.0	8.07	10.8	25.7
$\mathbf{E}[\nu_{sc} \vartheta \leq 1]$	3.20	3.45	7.23	97.3	7.50	7.47	17.0
$\mathbf{E}[\nu_{sc} \vartheta > 1]$	3.57	6.31	17.28	50.5	8.44	16.6	53.7
медиана ν_{sc}	3	3	6	36	6	7	12
99%-квантиль ν_{sc}	15	21	53	333	45	65	181
$\mathbf{P}(\nu_{un} \geq n^*)$	0.099	0.099	0.074	0.039	0.047	0.047	0.040
$\Psi(d_0)$	0.363	0.625	0.776	0.165	0.380	0.636	0.779

Выводы 2.2. Для гамма-показательной модели имеет место:

1. При незначительном превышении номинального уровня значимости последовательная процедура ν_{un} первого перескока с обязательной остановкой при достижении числа наблюдений $n = n^*$ в 2–4 раза понижает средний размер выборки.

2. Медиана момента остановки в 5–10 раз меньше НОВ.

3. Критерий, основанный на статистике вклада, также значительно сокращает средний объём проведенных испытаний.

4. Критерий, основанный на статистике вклада, имеет уровни d-рисков как значительные превышающие, так и значительно преуменьшающие заданные уровни надёжности.

2.1.3 Контроль по альтернативному признаку (модель В-В)

Задача проверки гипотез о вероятности какого-либо события — наиболее часто возникающая на практике статистическая задача. В качестве примера можно рассмотреть задачу аттестации партии кофе по относительной доле θ зёрен сорта «Арабика». Предположим, что по договору поставки в каждой партии эта доля не должна быть меньше θ_0 . Если учесть выгоду от продажи кофе и потери от штрафов за невыполнение условий на указанную долю, то производитель может позволить отпустить производителю не более $\beta_1 \cdot 100\%$ «некондиционных» партий кофе. Решение о кондиционности или некондиционности партии кофе принимается по результатам анализа части зёрен из партии. Так как партия обычно содержит очень большое количество зёрен, а контроль производится над её небольшой частью, то можно считать, что в эксперименте наблюдается выборка из распределения Бернулли с вероятностью успеха θ . Вполне уместно здесь предположить, что для каждой партии значение θ есть реализация случайной величины ϑ . Модель бета-распределения ϑ достаточно богата, поэтому представляется резонным выбрать эту модель в качестве априорного распределения для параметра ϑ .

В таблице 2.5 представлены результаты сравнения процедур при различных значениях параметров модели.

Таблица 2.5: Характеристики последовательных процедур; В-В модель

	$\beta_1 = \beta_0 = 0.10$				$\beta_0 = \beta_1 = 0.05$			
$(p; q)$	(0.1; 10)	(1; 1)	(5; 5)	(5; 5)	(0.1; 10)	(1; 1)	(5; 5)	(5; 5)
θ_0	0.01	0.5	0.5	0.4	0.01	0.5	0.5	0.4
n^*	31	15	91	71	153	63	379	303
Π_0	0.825	0.50	0.50	0.267	0.825	0.50	0.50	0.267
Процедура первого перескока, усечённая на уровне n^*								
$\mathbf{E} \tilde{\nu}_{un}$	16.9	6.8	33.0	19.9	53.8	16.9	90.1	65.6
Медиана	17	3	19	8	49	7	35	21
\mathcal{R}_0	0.100	0.109	0.119	0.119	0.064	0.061	0.066	0.065
\mathcal{R}_1	0.120	0.108	0.118	0.124	0.046	0.062	0.069	0.064
$\Psi(d_0)$	0.665	0.502	0.501	0.280	0.677	0.504	0.505	0.229
$\mathbf{P}(\nu_{un} \geq n^*)$	0.029	0.182	0.147	0.117	0.058	0.107	0.100	0.091
Процедура на статистике вклада								
$\mathbf{E} \nu_{sc}$	31.2	5.8	30.4	23.8	81.1	19.3	101.0	72.3
Медиана	33	3	21	15	73	11	60	40
\mathcal{R}_0	0.070	0.110	0.116	0.152	0.042	0.048	0.051	0.063
\mathcal{R}_1	0.136	0.107	0.115	0.080	0.057	0.048	0.053	0.046
$\Psi(d_0)$	0.867	0.50	0.50	0.243	0.852	0.50	0.501	0.249
$\mathbf{P}(\nu_{sc} \geq n^*)$	0.882	0.030	0.043	0.053	0.067	0.044	0.034	0.033

Выводы 2.3. Для модели бета-Бернулли справедливо следующее:

1. При незначительном превышении номинального уровня значимости момент остановки $\tilde{\nu}_{un}$ последовательной процедуры первого перескока с обязательной остановкой при достижении числа наблюдений $n = n^*$ и процедура на статистике вклада в 2–4 раза понижают средний размер выборки.

2. Процедура на статистике вклада за редким исключением «проигрывает» усечённой универсальной гарантийной процедуре, особенно с точки зрения медианы момента остановки.

3. При отклонении априорного среднего от границы проверяемых гипотез d-риски ПКСВ отличаются от номинальных уровней — d-риск принятия гипотезы с меньшей априорной вероятностью превышает заявленный уровень, d-риск противоположного решения меньше соответствующего уровня.

2.2 Введение в проблему множественного тестирования

D-апостериорный подход к проблеме гарантийности статистического вывода естественным образом может быть применён к решению задач множественного тестирования гипотез. Для подобного рода задач характерна ситуация, когда имеется реальная последовательность статистических экспериментов с различными случайными значениями тестируемого параметра, при этом каждый эксперимент должен заканчиваться принятием определенного решения, вероятность ошибочности которого должна быть гарантирована.

Обычно в таких задачах гарантируется «общий уровень значимости» типа FWER, PCER, FDR. Показатель FWER (family-wise error rate) равен вероятности отвержения хотя бы одной верной нулевой гипотезы. Другими словами, если имеется M статистических экспериментов, в каждом из которых проверяется некая нулевая гипотеза (возможно своя для каждого эксперимента) с помощью решающих функций $\delta_1, \dots, \delta_M$, то в предположении, что в экспериментах с номерами j_1, \dots, j_m справедлива нулевая гипотеза,

$$\text{FWER} = \mathbf{P} \left(\bigcup_{k=1}^m \{ \delta_{j_k} \neq d_0 \} \right).$$

Один из способов контроля FWER на уровне α основан на поправке Бонфер-

рони, в соответствии с которой в каждом из M статистических экспериментов уровень значимости выбирается равным α/M . Определения других, близких по смыслу к FWER, показателей общего уровня значимости и способы их контроля можно найти в обзоре [49, Sheffe], в главе 4 монографии [34, Efron] или в монографии [32, Dudoit, Laan].

В последние годы наиболее востребован показатель FDR (false discovery rate), предложенный в конце 90-х годов прошлого века Бенжамини и Хочбергом [28]. Этот показатель равен математическому ожиданию относительной доли ложно отвергнутых нулевых гипотез:

$$\text{FDR} = \text{FDR}(m) = \mathbf{E}_m \frac{V}{R},$$

где $R = \sum_{j=1}^M \mathbb{I}\{\delta_j \neq d_0\}$ — общее количество отвергнутых нулевых гипотез, $V = \sum_{j=1}^M \mathbb{I}\{\delta_j \neq d_0, H_j = H_0\}$ — количество ложно отвергнутых нулевых гипотез; математическое ожидание \mathbf{E}_m вычисляется в предположении, что истинное количество экспериментов со справедливыми нулевыми гипотезами равно m . Бенжамини и Хочберг предложили процедуру, использующую весь набор индивидуальных p -значений во всех M экспериментах, для которой $\text{FDR} \leq m\alpha/M$, где α — заданный уровень значимости. Точнее говоря, они использовали процедуру R. J. Simes [52], который показал ранее, что эта процедура гарантирует уровень α для показателя FWER и менее консервативна, чем процедуры, использующие поправку Бонферрони.

Средняя ошибка от принятия нулевой гипотезы (false negative rate) определяется аналогичным образом:

$$\text{FNR} = \mathbf{E}_m \frac{W}{M - R},$$

где $W = \sum_{j=1}^M \mathbb{I}\{\delta_j = d_0, H_j \neq H_0\}$ — количество ошибочно принятых нулевых гипотез. Этот показатель вовсе не поддается контролю вне рамок байесовского подхода.

Заметим, что при вычислении FDR отношение V/R по определению полагается равным нулю, если $R = 0$, т.е. отсутствуют эксперименты с отвергнутыми нулевыми гипотезами. Поскольку экспериментатора интересуют только ситуации, в которых $R > 0$, то вместо показателя FDR следовало бы рассмат-

ривать показатель

$$\text{pFDR} = \mathbf{E}_m(V/R | R > 0) \quad -$$

positive false discovery rate, т.е. условное математическое ожидание относительной доли ложно отвергнутых нулевых гипотез, если количество отвергнутых нулевых гипотез больше нуля. К сожалению, процедура Бенжамини–Хочберга не контролирует этот показатель. Было замечено (см. [60]), что в случае, когда количество нулевых гипотез m близко к общему количеству экспериментов M , процедура Бенжамини–Хочберга имеет высокую вероятность $\mathbf{P}(R = 0)$ и поэтому величина pFDR может значительно превышать заданный уровень α . С другой стороны, при достаточно большом количестве $M - m$ справедливых альтернатив, процедура Бенжамини–Хочберга излишне консервативна — величина pFDR существенно меньше α .

В дальнейшем в трудах Storey [55]– [57], эта характеристика была встроена в байесовскую парадигму. Предложенная им средняя ошибка pFDR (в байесовской постановке) полностью совпадает с понятием d-риска первого рода. Критическая область оптимальной процедуры, предложенной в [55], при попадании в которую нулевая гипотеза отвергается, совпадает с областью отклонения соответствующей гипотезы критерием вида (1.6). Таким образом, идеология d-апостериорного подхода может быть применена к задачам множественного тестирования с независимыми экспериментами.

В этой главе сначала оптимальный критерий (1.6) применяется к решению проблемы гарантийного выявления гиперактивных генов, ответственных за онкологическое заболевание (по данным из монографии Эфрона [34]). Затем предлагается общая байесовская модель для решения подобного рода задач с выявлением не только гиперактивных генов, но также и гипоактивных (подавленных) генов. Показывается, что при нормальном распределении данных и нормальном распределении тестируемого параметра критерий Стьюдента, обычно используемый для решения подобных задач, не является оптимальным. Оптимальный критерий, минимизирующий d-риск второго рода должен быть основан на специфической линейной комбинации выборочных средних.

Суть нашего взгляда на проблему множественного тестирования, как и методы решения этой проблемы, поясним на следующем примере из области

генетики. Рассмотрим часто решаемую задачу выявления генов, ответственных за определенную патологию, например, за онкологическое заболевание (назовём их «плохими»). Обычно плохие гены обладают повышенной экспрессией — гиперактивные гены. Гены с нормальной экспрессией будем называть «хорошими». Выявление генов с изменённым показателем экспрессии обычно осуществляется по данным замеров экспрессии у большого числа генов с последующим применением методов проверки статистических гипотез к каждому из них.

Отбирается большое число $N \gg 1$ генов, которые согласно генетическим показаниям могут быть ответственными за появление болезни или, что чаще, гены выбираются случайным образом. Экспрессия отобранных генов измеряется у пациентов двух групп: контрольной группы, состоящей из n_1 здоровых индивидуумов, и группы n_2 пациентов, страдающих заболеванием. Для каждого из N генов по результатам замеров проверяется гипотеза однородности распределений в двух группах с помощью одного из стандартных двухвыборочных критериев (Стьюдента, Вилкоксона и т.п.). При этом, памятуя о множественности принимаемых решений, уровень значимости этого критерия стараются выбирать так, чтобы гарантировать некую общую характеристику надёжности тестирования всей совокупности генов.

Как отмечено выше, обычно эта ситуация рассматривается как проблема множественного тестирования и для её решения применяется совокупность процедур, гарантирующая «общий уровень значимости» типа FWER, PCER и т.п. Несоответствие методов множественного тестирования задачам, названным в последствии Large-scale inference, было понято давно. В основном акцент делался на изучение доли хороших генов, среди объявленных плохими. Это объясняется, конечно, тем, что при этом часто не требуется построения вероятностной модели и достаточно управлять только классической вероятностью ошибки первого рода. Так в работе Sorić [53] (где можно найти упоминания и о более ранних работах в этом направлении) строятся толерантные границы для этой доли.

Как нам кажется, в таких исследованиях генетиков должны интересоваться совсем другие характеристики надёжности статистического вывода. Здесь воз-

можны две ситуации. Если после проведения анализа гены, объявленные «плохими», будут подвергаться дальнейшему изучению, например, с целью выявления причин их повышенной активности, то исследователю важно иметь среди таких генов мало в действительности «хороших». Другими словами, доля «хороших», среди объявленных «плохими», должна быть мала. В ситуации же, когда гены, объявленные «плохими», будут подвергаться воздействию, например, в надежде на излечение от болезни, то исследователю важно добиться уменьшения доли «плохих» генов среди объявленных «хорошими».

Аналогичные ситуации возникают и при статистическом контроле качества массовой продукции, когда обычно контролируется вероятность ошибки первого рода, хотя по сути, необходимо контролировать долю плохой продукции среди успешно прошедшей контроль (отправленной потребителю).

2.2.1 Тестирование экспрессии генов

Методы d-апостериорного подхода применяются к задаче выявления различий между экспрессией генов в двух группах пациентов на примере данных, взятых из гл. I монографии Эфрона [34]. Данные, на которых проводится статистический анализ, представляют собой измерения показателя экспрессии $M = 6033$ генов у каждого из $n = 102$ пациентов; у $n_2 = 52$ пациентов диагностировано заболевание, $n_1 = 50$ пациентов признаны здоровыми. Для каждого из генов вычисляется стандартная двухвыборочная статистика Стьюдента T с $n_1 + n_2 - 2 = 100$ степенями свободы. В Таблице 2.6 приведены частоты попадания значений статистик Стьюдента в интервалы гистограммы, которая приведена на стр. 17 монографии Эфрона [34]. На этой гистограмме имеется 96 интервалов от -4.8 до $+4.8$ с шагом 0.1. В таблице соответствующие этим интервалам частоты расположены построчно.

Если соблюдены условия применения двухвыборочного критерия Стьюдента, то статистика T имеет нецентральное распределение Стьюдента с параметром нецентральности

$$\delta_n = \theta\gamma_n, \quad \gamma_n = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}, \quad (2.1)$$

Таблица 2.6: Данные по экспрессии генов из монографии Б. Эфрона

0	0	0	1	1	1	1	1	1	4	1	4	5	4	6	4
1	5	9	6	10	12	14	17	23	23	28	31	39	35	52	71
71	82	90	118	135	153	152	179	202	185	178	203	205	205	233	204
188	227	210	230	214	188	218	165	159	146	123	139	106	109	90	82
67	63	52	33	28	28	28	19	17	12	9	11	9	9	4	9
9	2	3	3	3	1	1	2	1	2	2	1	1	1	2	2

где θ — нормированная на стандартное отклонение разность средних значений экспрессий у плохих и хороших генов. Хорошо известно, что при столь большом числе степеней свободы (100) распределение статистики T практически неотличимо от нормального $\mathcal{N}(\delta_n, 1)$ закона.

Таким образом, имеем $M = 6033$ экспериментов, в каждом из которых на основе реализации случайной величины $T \sim \mathcal{N}(\theta\gamma_n, 1)$ необходимо принять или отвергнуть гипотезу $H_0 : \theta = 0$ (или $H_0 : \theta \leq 0$) при альтернативе $H_1 : \theta \neq 0$ ($H_1 : \theta > 0$). Характеристика статистической процедуры, связанная с долей плохих генов среди тех, что объявлены хорошими, может быть вычислена только, если допустимо предположение, что значения параметра $\theta_1, \dots, \theta_M$ в каждом эксперименте есть реализации некоторой случайной величины (обозначим ее ϑ).

В монографии Б. Эфрона [34] рассматривается модель, в которой ϑ может принимать только два значения: $\theta_0 = 0$ с вероятностью π_0 или θ_1 с вероятностью $\pi_1 = 1 - \pi_0$ ($\theta_1 \neq 0$ — некоторое неизвестное значение). В рамках этой модели последовательность T_1, \dots, T_M есть последовательность случайных величин, имеющих функцию распределения и плотность вида

$$F(t) = \pi_0\Phi(t) + \pi_1\Phi(t - \theta_1), \quad f(t) = \pi_0\varphi(t) + \pi_1\varphi(t - \theta_1),$$

где Φ , φ — функция распределения и плотность нормального закона со средним 0 и единичной дисперсией. Хотя в монографии Б. Эфрона предполагается нормальность только нулевого распределения, но, как было сказано выше, и альтернативное нецентральное распределение Стюдента с достаточной для практических целей точностью также приблизительно нормально, но с другим средним.

Представленная в монографии (рис. 2.1, стр. 17) гистограмма 6033 наблюденных значений статистики T плохо согласуется с приведённой выше моделью смеси двух нормальных распределений: p -значение критерия минимума хи-квадрат меньше 10^{-11} . Следует также отметить, что данные имеют почти нулевой выборочный коэффициент асимметрии. По всей видимости, это указывает на присутствие в данных не только гиперактивных, но и гипоактивных (подавленных) генов.

2.2.2 Вероятностная модель с тремя параметрами

Мы предлагаем следующее видоизменение модели Эфрона. Заметим, что эта модель возникла с подачи специалистов-биологов, считающих, что уровень экспрессии в группе больных для каждого гена различен и может быть как повышенным, так и пониженным. Безусловная (априорная) функция распределения ϑ в этой модели представима в виде

$$\mathbf{P}(\vartheta < \theta) = (1 - \pi)\mathbb{I}_0(\theta) + \pi G(\theta), \quad (2.2)$$

где $\mathbb{I}_0(\theta)$ — индикаторная функция множества $\theta \in (0, \infty)$, $G(\theta)$ — некоторая функция распределения, π ($= \pi_1$) — доля генов с изменённым уровнем экспрессии. Плотность безусловного распределения статистики T для этой модели равна

$$f_n(t) = (1 - \pi)\varphi(t) + \pi \int_{-\infty}^{\infty} \varphi(t - \theta\gamma_n) dG(\theta).$$

Исследуем поведение характеристик статистических процедур для двух параметрических моделей G . В первой из этих моделей функция $G(\theta) = \Phi((\theta - \mu)/\tau)$, то есть совпадает с функцией распределения нормального $\mathcal{N}(\mu, \tau^2)$ закона. Легко понять, что в этом случае безусловная функция плотности и безусловная функция распределения T равны, соответственно,

$$\begin{aligned} f(t; \pi, \mu, \tau, \gamma_n) &= (1 - \pi)\phi(t) + \pi \frac{1}{\sqrt{1 + \gamma_n^2\tau^2}} \phi\left(\frac{t - \mu\gamma_n}{\sqrt{1 + \gamma_n^2\tau^2}}\right), \\ F(t; \pi, \mu, \tau, \gamma_n) &= (1 - \pi)\Phi(t) + \pi \Phi\left(\frac{t - \mu\gamma_n}{\sqrt{1 + \gamma_n^2\tau^2}}\right), \quad t \in \mathbb{R}^1. \end{aligned} \quad (2.3)$$

Проверка согласия этой модели с описанными выше гистограммными данными ($\gamma_n = \sqrt{50 * 52 / (50 + 52)} \approx 5.05$) по методу минимума хи-квадрат показала прекрасное совпадение: p -значение 0.78 при 92 степенях свободы. Метод максимального правдоподобия дал следующие оценки параметров:

$$\tilde{\pi} = 0.149, \quad \tilde{\mu} = 0.016, \quad \tilde{\tau} = 0.271. \quad (2.4)$$

Заметим, что здесь можно выбрать $\tilde{\mu} = 0$, т.к., во-первых, это практически не влияет на значения d -рисков процедур. Во-вторых, в симметричной ситуации с $\mu = 0$ можно установить ряд интересных фактов относительно свойств рассматриваемых процедур. Кроме того, p -значение критерия хи-квадрат при такой оценке μ также больше 0.70.

Двусторонняя альтернатива. Рассмотрим задачу выделения генов с изменённой (в ту или иную сторону) экспрессией. Другими словами, для каждого гена необходимо проверить гипотезу $H_0 : \theta = 0$ при альтернативе $H_1 : \theta \neq 0$. Для построения оптимального критерия (1.6) необходимо вычислить апостериорную вероятность справедливости нулевой гипотезы:

$$\Pi_0(t) = \mathbf{P}(\vartheta = 0 | t) = \frac{(1 - \pi) \phi(t)}{f(t; \pi, \mu, \tau, \gamma_n)}.$$

Теорема 2.1. Пусть справедлива модель (2.2) с нормальным распределением G . 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 $C \leq \Pi_0(t^*)$) выполняется тогда и только тогда, когда $t^* - c \leq t \leq t^* + c$, где $c \geq 0$ и $\Pi_0(t^* - c) = \Pi_0(t^* + c) = C$.

Доказательство. Функция $1/\Pi_0(t)$ зависит от t через отношение стандартных нормальных плотностей $\phi((t - \mu\gamma_n)/\sqrt{1 + \tau^2\gamma_n^2})/\phi(t)$. Полагая $t = y - \mu/(\gamma_n\tau^2)$, получаем

$$\frac{\phi((t - \mu\gamma_n)/\sqrt{1 + \tau^2\gamma_n^2})}{\phi(t)} \Big|_{t=y-\mu/(\gamma_n\tau^2)} = \exp \left\{ \frac{y^2\tau^4\gamma_n^2 - \mu^2(1 + \tau^2\gamma_n^2)}{2\tau^2(1 + \tau^2\gamma_n^2)} \right\}.$$

Таблица 2.7: d-Риски и число решений d_1 критерия $|T| > c$ при различении гипотез $H_0 : \theta = 0$, $H_1 : \theta \neq 0$; модель с 3 параметрами

	$c = 1.409$	$c = 3.056$	$c = 3.194$	$c = 3.510$	$c = 4.126$
$\mathcal{R}_1(c)$	0.689	0.131	0.10	0.05	0.01
$\mathcal{R}_0(c)$	0.10	0.131	0.133	0.136	0.139
$\#(d_1)$	1169	80	70	42	15
$\text{ВН}(d_1)^*$	591	57	41	20	0

*) В строке $\text{ВН}(d_1)$ приведено количество отвержений нулевой гипотезы процедурой Бенжамини–Хочберга при ограничениях на величину FDR, совпадающих с величиной \mathcal{R}_1 .

В правой части здесь находится чётная функция аргумента y . Более того, эта функция убывает при $y < 0$ и возрастает при $y > 0$.

Утверждение (iii) очевидным образом следует из первых двух утверждений леммы и из непрерывности апостериорной вероятности по значению статистики $T = t$. \square

Таким образом, критерий, основанный на статистике Стьюдента T , должен отвергать гипотезу $H_0 : \theta = 0$, когда $|T + \mu/(\gamma_n\tau^2)| \geq c$, где константа c находится из условий на d-риск 1-го рода:

$$\begin{aligned} \mathcal{R}_1(c) &= \mathbf{P}(\vartheta = 0 \mid |T - t^*| \geq c) = \\ &= \frac{(1 - \pi)(1 - \Phi(t^* + c) + \Phi(t^* - c))}{1 - F(t^* + c; \pi, \mu, \tau, \gamma_n) + F(t^* - c; \pi, \mu, \tau, \gamma_n)} = \beta_1 \end{aligned} \quad (2.5)$$

или на d-риск 2-го рода:

$$\begin{aligned} \mathcal{R}_0(c) &= \mathbf{P}(\vartheta \neq 0 \mid |T - t^*| < c) = \\ &= 1 - \frac{(1 - \pi)(\Phi(t^* + c) - \Phi(t^* - c))}{F(t^* + c; \pi, \mu, \tau, \gamma_n) - F(t^* - c; \pi, \mu, \tau, \gamma_n)} = \beta_0. \end{aligned} \quad (2.6)$$

Положим в этих уравнениях параметры, равные оценкам (2.4), и множитель $\gamma_n = 5.05$. В таблице 2.7 приведены значения критических констант, найденных из уравнений (2.5), (2.6). Кроме того, эта таблица содержит данные о количестве $\#(d_1)$ отклонений нулевой гипотезы при соответствующих критических константах. Для сравнения здесь указаны также значения числа отклонений нулевой гипотезы в результате применения процедуры Бенжамини–

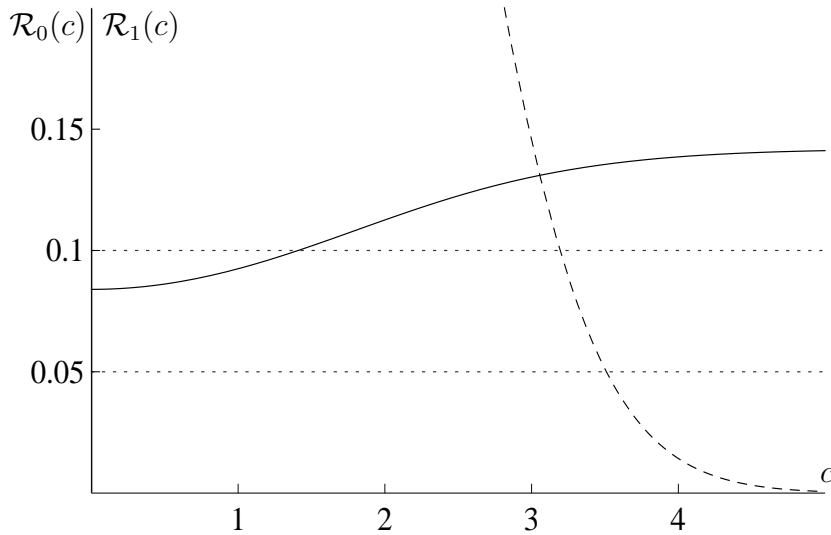


Рис. 2.1: Изменение d-рисков в зависимости от критической константы

Хочберга, в которой уровень значимости выбран совпадающим с величиной соответствующего d-риска \mathcal{R}_1 .

Добиться ограничений на величину d-риска $\mathcal{R}_0(c)$ можно не для всех значений β_0 (см. рисунок 2.1, где сплошная линия показывает график изменения $\mathcal{R}_0(c)$, а пунктирная — график $\mathcal{R}_1(c)$). Связано это со следующим обстоятельством. По хорошо известному телескопическому свойству условного математического ожидания d-риск есть условное математическое ожидание от апостериорной вероятности справедливости альтернативной гипотезы. Поэтому d-риск любой статистической процедуры не может быть меньше минимального (относительно результатов наблюдений) значения соответствующей апостериорной вероятности. В данной модели при указанных значениях параметров и альтернативе $H_1 : \theta \neq 0$ имеем

$$\mathcal{R}_0(c) \geq \min_t \mathbf{P}(\vartheta \neq 0 \mid |T| = t) = \mathbf{P}(\vartheta \neq 0 \mid |T| = 0) = 0.084$$

для любой критической константы c . Максимальные значения d-рисков совпадают с априорными вероятностями соответствующих альтернативных гипотез. Таким образом, для параметров модели (2.4) при любых $c \geq 0$ справедливо

$$0.084 \leq \mathcal{R}_0(c) \leq \mathbf{P}(\vartheta \neq 0) = 0.141, \quad 0 \leq \mathcal{R}_1(c) \leq 0.859.$$

Если требуется построить критерий с заданными ограничениями на величину обоих d-рисков \mathcal{R}_0 и \mathcal{R}_1 , то необходимо соответствующим образом

подобрать объём выборочных данных (количество обследованных пациентов). Свойства критерия существенным образом зависят от коэффициента γ_n .

Выводы 2.4. Для задачи выделения генов с изменённой экспрессией в модели с тремя параметрами справедливы следующие утверждения.

1. Процедура Бенжамини–Хочберга более консервативна, чем оптимальный d -гарантийный тест — она почти вдвое реже отвергает нулевую гипотезу.

2. При увеличении объёма каждой из выборок в 5 раз (при этом γ_n^2 также увеличивается в 5 раз) d -риск $\mathcal{R}_0(0.89) = 0.05$, а d -риск $\mathcal{R}_1(0.89) = 0.74$.

3. Если же γ_n^2 увеличить в 20 раз, то при выборе критической константы $C = 2.757$ рассматриваемый критерий будет гарантийным в d -апостериорном смысле: $\mathcal{R}_0(2.757) = \mathcal{R}_1(2.757) = 0.0494$.

Односторонние гипотезы. В монографии Б. Эфрона относительно представленных данных проверяется односторонняя гипотеза $H_0 : \theta \leq 0$ при альтернативе $H_1 : \theta > 0$. В этой задаче ввиду монотонности отношения правдоподобия оптимальный d -гарантийный критерий отвергает гипотезу при значениях статистики T , больших некоторой константы c . Риски такого критерия вычисляются по следующим формулам:

$$\begin{aligned} \mathcal{R}_0(c) &= \mathbf{P}(\vartheta > 0 \mid T \leq c) = \frac{\pi}{F(c; \pi, \mu, \tau, \gamma_n)} \cdot \int_{-\mu/\tau}^{\infty} \Phi(c - \gamma_n \mu - \gamma_n \tau y) \phi(y) dy, \\ \mathcal{R}_1(c) &= \mathbf{P}(\vartheta \leq 0 \mid T > c) \\ &= \frac{1}{1 - F(c; \pi, \mu, \tau, \gamma_n)} \cdot \left[(1 - \pi) \Phi(-c) + \pi \int_{-\infty}^{-\mu/\tau} \Phi(\gamma_n \tau y + \gamma_n \mu - c) \phi(y) dy \right]. \end{aligned} \tag{2.7}$$

где $F(c; \pi, \mu, \tau, \gamma_n)$ — вероятность принятия нулевой гипотезы $\mathbf{P}(d_0)$.

Таблица 2.8 содержит значения критических констант c при различных ограничениях на d -риски, а также данные о количестве $\#(d_1)$ принятых решений d_1 (среди всех 6033 генов) об ответственности гена за онкологию при использовании критической области $T > c$, предназначенной для тестирования гипотезы $H_0 : \theta \leq 0$ против альтернативы $H_1 : \theta > 0$. Заметим, что для

Таблица 2.8: d-Риски и число решений d_1 критерия $T > c$ при различении гипотез $H_0 : \theta \leq 0$, $H_1 : \theta > 0$; модель с 3 параметрами

	$c = -1.753$	$c = 1.390$	$c = 3.199$	$c = 3.387$	$c = 3.514$	$c = 4.130$
$\mathcal{R}_1(c)$	0.925	0.624	0.10	0.067	0.05	0.01
$\mathcal{R}_0(c)$	0.01	0.05	0.066	0.067	0.068	0.069
$\#(d_1)$	5746	367	36	25	22	11
$\text{ВН}(d_1)^*)$	2403	219	25	19	15	5

*) $\text{ВН}(d_1)$ — количество отвержений нулевой гипотезы процедурой Бенжамини–Хочберга.

данной задачи области возможных значений d-рисков задаются неравенствами

$$0 \leq \mathcal{R}_0(c) \leq \mathbf{P}(\vartheta > 0) = 0.071, \quad 0 \leq \mathcal{R}_1(c) \leq 0.929.$$

Как видно из таблицы, здесь можно построить критерий, обладающий приемлемыми средними ошибками обоих типов.

Выводы 2.5. Для задачи выделения генов с повышенной экспрессией в модели с тремя параметрами справедливы следующие утверждения.

1. Процедура Бенжамини–Хочберга более консервативна, чем оптимальный d-гарантийный тест — она в полтора-два раза реже отвергает нулевую гипотезу.

2. Если количество пациентов увеличить так, чтобы коэффициент $\gamma_n^2 = n_1 n_2 / (n_1 + n_2)$ увеличился в 3.5 раза с 25.5 для рассматриваемого примера до 90, то можно построить критерий, у которого оба d-риска \mathcal{R}_0 , \mathcal{R}_1 не превосходят $\beta = 0.05$.

Различение трёх гипотез. Необходимость отдельного анализа гипоактивных и гиперактивных генов приводит к необходимости построения критерия различения трёх гипотез относительно значения разности экспрессии в группах пациентов

$$H_- : \theta < 0, \quad H_0 : \theta = 0, \quad H_+ : \theta > 0.$$

Поскольку проблема оптимального различения трёх и более гипотез в рамках d-апостериорного подхода пока не решена, обратимся к традиционному использованию критерия Стьюдента: гипотеза H_0 принимается, если статистика

Стьюдента $-a < T < b$, в противном случае принимается соответствующая альтернатива. Обозначим критерий подобного типа через $\delta_{a,b}$. Рассмотрим три характеристики этого критерия, интерпретируемые как надёжность принятых решений:

$$\begin{aligned} Q_0(a, b) &= \mathbf{P}(\vartheta = 0 \mid -a < T < b), \\ Q_-(a) &= \mathbf{P}(\vartheta < 0 \mid T \leq -a), \\ Q_+(b) &= \mathbf{P}(\vartheta > 0 \mid T \geq b). \end{aligned} \tag{2.8}$$

Эти характеристики, как и величины d-рисков, имеют следующие границы изменений для любых критических констант a, b :

$$\begin{aligned} \mathbf{P}(\vartheta = 0) &\leq Q_0(a, b) \leq \sup_t \mathbf{P}(\vartheta = 0 \mid T = t) = 0.917, \\ \mathbf{P}(\vartheta < 0) &\leq Q_-(a) \leq \sup_t \mathbf{P}(\vartheta < 0 \mid T = t) = 1, \\ \mathbf{P}(\vartheta > 0) &\leq Q_+(b) \leq \sup_t \mathbf{P}(\vartheta > 0 \mid T = t) = 1, \end{aligned} \tag{2.9}$$

где максимальные значения апостериорных вероятностей вычислены для приведённых в (2.4) оценок параметров.

Замечание 15. Есть некоторые основания полагать, что в случае, когда априорное среднее $\mu = 0$ среди подобных критериев наилучшим (минимаксным) будет критерий $\delta_{c,c}^*$, у которого $Q_0(c, c) = Q_-(c) = Q_+(c)$. Для параметров $\mu = 0$, $\pi = 0.141$, $\tau = 0.30$ значение критической константы такого критерия $c = 3.068$ и

$$Q_0(c, c) = Q_-(c) = Q_+(c) = 0.870.$$

Для несимметричной ситуации, когда $\mu \neq 0$, скажем, $\mu = 1$, минимаксным будет критерий также с совпадающими значениями надёжности

$$Q_-(a) = Q_0(a, b) = Q_+(b) = 0.881$$

при $a = 3.733$, $b = 2.660$. Показать минимаксность этих критериев нам удалось только численными методами.

Обсудим возможность построения гарантийного (в каком-либо смысле) критерия различения трёх гипотез. Как и в случае с двумя гипотезами, критерий, удовлетворяющий ограничениям на надёжность нескольких принятых решений, существует только при соответствующем значении $\gamma_n^2 = n_1 n_2 / (n_1 + n_2)$.

Таблица 2.9: Надёжность и число решений d_{\pm} симметричного критерия при различении гипотез $H_- : \theta < 0$, $H_0 : \theta = 0$, $H_+ : \theta > 0$; модель с 3 параметрами

	$c = 1.421$	$c = 2.679$	$c = 2.992$	$c = 3.202$	$c = 3.517$	$c = 4.132$
$Q_0(c)$	0.90	0.875	0.871	0.868	0.865	0.862
$Q_-(c)$	0.30	0.75	0.85	0.90	0.95	0.99
$Q_+(c)$	0.30	0.75	0.85	0.90	0.95	0.99
$\#(d_{\pm})$	1169	143	89	70	42	15

Для данных рассматриваемого здесь примера из монографии Б. Эфрона $\gamma_n^2 = n_1 n_2 / (n_1 + n_2) \approx 25.5$ с $n_1 = 50$, $n_2 = 52$.

В таблице 2.9 приведены значения критических констант, соответствующих различным ограничениям на величины d-рисков с модельными параметрами (2.4). Кроме того, здесь приведены данные о количестве $\#(d_{\pm})$ принятых решений в пользу одной из альтернатив $H_- : \theta < 0$ или $H_+ : \theta > 0$.

Выводы 2.6. Для задачи выделения генов с повышенной, пониженной или неизменённой экспрессией в модели с тремя параметрами справедливы следующие утверждения.

1. Для имеющихся в наличии групп испытуемых пациентов с $\gamma_n^2 \approx 25.5$ в силу ограничений (2.9) нельзя добиться надёжности решения в пользу нулевой гипотезы $H_0 : \theta = 0$ выше, чем 0.917. Однако надёжность любого другого решения можно сделать сколь угодно высокой.

2. Увеличив количество испытуемых в каждой группе чуть больше чем в 3 раза (точнее, при $\gamma_n^2 \approx 78$), можно обеспечить надёжность всех трёх принимаемых решений на уровне 90% :

$$Q_-(c) = Q_0(c, c) = Q_+(c) = 0.90$$

для $c = 2.752$.

3. Надёжности на уровне 95% для всех трёх решений можно достичь только при $\gamma_n^2 \approx 498$ — примерно по тысяче пациентов в каждой группе.

2.2.3 Модель с шестью параметрами

Разделение генов на гипоактивные (с пониженной экспрессией) и гиперак-

тивные (с повышенной экспрессией) среди заболевших пациентов можно описать с помощью следующего уточнения вероятностной модели (2.2). В этой модели априорное распределение ϑ разделяется на три компоненты. Любой ген с вероятностью π_1 может быть гипоактивным и разность средних значений θ в двух группах для этого гена есть реализация случайной величины $\vartheta_1 \sim \mathcal{N}(\mu_1, \tau_1^2)$. Аналогично для гиперактивных генов имеем с вероятностью π_2 случайную величину $\vartheta_2 \sim \mathcal{N}(\mu_2, \tau_2^2)$. Следовательно, доля «хороших» генов равна $1 - \pi_1 - \pi_2$ и для них разность средних равна нулю. Маргинальная функция распределения наблюдаемой статистики в новой модели имеет вид

$$F(t; \zeta) = (1 - \pi_1 - \pi_2)\Phi(t) + \pi_1 \Phi\left(\frac{t - \mu_1\gamma_n}{\sqrt{1 + \tau_1^2\gamma_n^2}}\right) + \pi_2 \Phi\left(\frac{t - \mu_2\gamma_n}{\sqrt{1 + \tau_2^2\gamma_n^2}}\right),$$

где вектор шести параметров $\zeta = (\pi_1, \mu_1, \tau_1, \pi_2, \mu_2, \tau_2)$ и, как и выше, $\gamma_n^2 = n_1n_2/(n_1 + n_2)$. Безусловная плотность $f(t; \zeta)$ имеет похожую структуру.

Согласие данных с этой моделью столь же высокое: p -значение критерия минимума хи-квадрат ≈ 0.77 . Оценки параметров по методу минимума хи-квадрат

$$\begin{aligned} \hat{\pi}_1 &= 0.0244, & \hat{\mu}_1 &= -0.4632, & \hat{\tau}_1 &= 0.0418, \\ \hat{\pi}_2 &= 0.0370, & \hat{\mu}_2 &= 0.3792, & \hat{\tau}_2 &= 0.1642. \end{aligned} \tag{2.10}$$

Двусторонняя альтернатива. Рассмотрим задачу выделения генов с изменённой экспрессией, т.е. задачу различения гипотез $H_0 : \theta = 0$ и $H_1 : \theta \neq 0$, при значениях параметров (2.10) и $\gamma_n = \sqrt{n_1n_2/(n_1 + n_2)} = 5.05$. При этих параметрах минимальное значение апостериорной вероятности справедливости нулевой гипотезы $\mathbf{P}(\vartheta = 0 | T = t)$ равно 0, а максимальное — 0.989. Поэтому для любого $\beta_1 \leq 0.989$ здесь можно построить критерий δ , у которого d-риск $\mathcal{R}_1(\delta) \leq \beta_1$. Ограничение на величину d-риска $\mathcal{R}_0(\delta) \leq \beta_0$ достижимо только для $\beta_0 \geq 0.011$. Так как $\mathcal{R}_0(\delta) \leq \mathbf{P}(\vartheta = 0) = 0.0614$, то для $\beta_0 \geq 0.0614$ оптимальный критерий не строится.

Апостериорная вероятность $\Pi_0(t) = \mathbf{P}(\vartheta = 0 | T = t)$ в данном случае представляет собой непрерывную функцию аргумента t с единственным

Таблица 2.10: d-Риски и число решений d_1 критерия $|T| > c$ при различении гипотез $H_0 : \theta = 0$, $H_1 : \theta \neq 0$; модель с 6 параметрами

	$c = 0.821$	$c = 1.995$	$c = 3.110$	$c = 3.231$	$c = 3.559$	$c = 4.231$
$\mathcal{R}_1(c)$	0.879	0.583	0.127	0.10	0.05	0.01
$\mathcal{R}_0(c)$	0.015	0.03	0.05	0.052	0.055	0.059
$\#(d_1)$	2778	432	80	70	34	12
$VH(d_1)$	3549	373	57	42	20	0

*) $VH(d_1)$ — количество отвержений нулевой гипотезы процедурой Бенжамини–Хочберга.

максимумом в точке $t < 0$. Поэтому область принятия нулевой гипотезы оптимального критерия (1.6) может быть представлена в виде $c_1 < T < c_2$, где $\Pi_0(c_1) = \Pi_0(c_2)$. К сожалению, функция $\Pi_0(t)$ не обладает симметрией по переменной t , что несколько осложняет процесс построения критерия (сравните с ситуацией для предыдущей модели). С другой стороны, влияние этой несимметричности на величину d-рисков настолько мало, что ею можно пренебречь. Поэтому в таблице 2.10 мы ограничились рассмотрением критериев, принимающих нулевую гипотезу, если $|T| < c$, $c > 0$.

Различение трёх гипотез. Рассмотрим теперь задачу различения трёх гипотез $H_0 : \theta = 0$, $H_- : \theta < 0$, $H_+ : \theta > 0$. Снова остановимся на рассмотрении критериев, принимающих гипотезу H_0 , если статистика Стьюдента $-c_1 < T < c_2$, в противном случае принимается соответствующая альтернатива. Качество этих критериев оценивается тремя величинами (2.8), интерпретируемыми как надёжность принятых решений.

Отметим, что при значениях параметров, приведённых в (2.10), максимум апостериорной вероятности $\Pi_0(t) = \mathbf{P}(\vartheta = 0 | T = t)$ равен 0.989 и достигается при $t = -0.277$ (не в нуле!). Поэтому здесь можно подобрать константы c_1, c_2 так, чтобы выполнялось ограничение $Q_0(c_1, c_2) \geq 0.95$ и, даже, $Q_0(c_1, c_2) \geq 0.975$.

Замечание 16. Равенство $Q_0(c_1, c_2) = \beta$ может быть достигнуто при многих значениях c_1, c_2 . Есть некоторые основания надеяться, что при выборе этих констант так, что апостериорная вероятность $\Pi_0(-c_1) = \Pi_0(c_2)$, то мини-

Таблица 2.11: Надёжность и число решений d_{\pm} несимметричного критерия при различении гипотез $H_- : \theta < 0$, $H_0 : \theta = 0$, $H_+ : \theta > 0$; модель с 6 параметрами

$-c_1$	-3.284	-3.632	-3.588	-4.375	-3.632	-2.787
c_2	3.188	3.503	3.464	4.142	2.746	3.503
Q_-	0.90	0.95	0.945	0.99	0.95	0.763
Q_+	0.90	0.95	0.945	0.99	0.773	0.95
Q_0	0.948	0.945	0.945	0.941	0.95	0.95
$\#(d_{\pm})$	66	37	37	12	93	77
$\text{ВН}(d_1)^*)$	42	20	25	0	20	20

*) $\text{ВН}(d_1)$ — количество отвержений нулевой гипотезы процедурой Бенжамини–Хочберга.

мум надёжности двух других решений $\min\{Q_-(c_1), Q_+(c_1)\}$ будет максимальным. Например, для рассматриваемого здесь примера при $c_1 = 3.118$, $c_2 = 3.102$ апостериорная вероятность $\Pi_0(-c_1) = \Pi_0(c_2)$, при этом надёжность $Q_0(c_1, c_2) = 0.95$, а $Q_-(c_1) = 0.864$, $Q_+(c_2) = 0.881$. Если же критические константы выбирать симметрично: $c_1 = c_2 = 3.110$, то имеем $Q_0(c_1, c_2) = 0.95$ и $Q_-(c_1) = 0.861$, $Q_+(c_2) = 0.883$.

2.3 Общая модель сравнения двух групп

Поскольку измерения экспрессии генов производятся в двух различных группах, проще всего рассматривать эту ситуацию как наблюдения над реализациями двух случайных величин X (в контрольной группе) и Y (в группе больных). Предполагается, что плотности распределений X и Y принадлежат некоторому параметрическому семейству $\{f(x|t), t \in \mathcal{T} \subset R\}$. Для каждого изучаемого гена параметр t у распределений X и Y одинаков, если ген не ответственен за патологию, и может быть разным в противном случае: t для контрольной группы и $t + \theta$ для группы с заболеванием. По результатам наблюдений за экспрессией данного гена у пациентов в двух группах требуется проверить одну из гипотез вида $H_1 : \theta \neq 0$ (ген с изменённой экспрессией), $H_1 : \theta > 0$ (ген гиперактивен) или $H_1 : \theta < 0$ (ген гипоактивен) при соответствующей противоположной альтернативе H_0 . Такого рода гипотезы проверяются по наблюдениям каждого гена и составная процедура принятия

решений по всей совокупности генов будет строиться в рамках байесовской модели априорных распределений t и θ .

Значения параметров t и θ для каждого гена трактуются как реализации соответствующих компонент случайного вектора (τ, ϑ) . Пусть $g_1(t)$ — априорная функция плотности τ . Значение параметра ϑ с вероятностью $(1 - \pi)$ равно 0 и с вероятностью π есть реализация случайной величины с условной относительно τ функцией плотности $g_2(\theta | \tau)$. (Плотность g_2 может быть смесью нескольких, например, двух плотностей, соответствующих значениям параметра θ у гипоактивных и гиперактивных генов.)

Пусть теперь $\mathbf{X} = (X_1, \dots, X_n)$ случайная выборка, описывающая статистический эксперимент по наблюдению экспрессии некоторого фиксированного гена в контрольной группе, $\mathbf{Y} = (Y_1, \dots, Y_m)$ — случайная выборка, соответствующая наблюдениям того же гена в группе заболевших пациентов. Плотность вектора (\mathbf{X}, \mathbf{Y}) при фиксированных значениях параметров $(\tau, \vartheta) = (t, \theta)$ равна

$$f(\mathbf{x}, \mathbf{y} | t, \theta) = f(\mathbf{x} | t)f(\mathbf{y} | t + \theta),$$

где $\mathbf{x} = (x_1, \dots, x_n)$, $\mathbf{y} = (y_1, \dots, y_m)$,

$$f(\mathbf{x} | t) = \prod_{i=1}^n f(x_i | t), \quad f(\mathbf{y} | t + \theta) = \prod_{j=1}^m f(y_j | t + \theta).$$

Совокупность измерений экспрессии всех N генов может быть описана теперь как совокупность N реализаций случайного вектора (\mathbf{X}, \mathbf{Y}) с безусловной (маргинальной) функцией плотности

$$\begin{aligned} f(\mathbf{x}, \mathbf{y}) &= (1 - \pi) \int_R f(\mathbf{x} | t)f(\mathbf{y} | t)g_1(t) dt \\ &\quad + \pi \int_{R^2} f(\mathbf{x} | t)f(\mathbf{y} | t + \theta)g_2(\theta | t)g_1(t) d\theta dt \\ &= \int_R f(\mathbf{x} | t)f_\pi(\mathbf{y} | t) g_1(t) dt, \end{aligned}$$

где $f_\pi(\mathbf{y} | t) = (1 - \pi)f(\mathbf{y} | t) + \pi \int_R f(\mathbf{y} | t + \theta)g_2(\theta | t) d\theta$.

Оптимальному критерию различения гипотез $H_0 : \theta \in \Theta_0$ и $H_1 : \theta \notin \Theta_0$ с ограничениями на d -риски соответствует тестовая статистика, равная

апостериорной вероятности справедливости H_0 ([21]):

$$\begin{aligned} \mathcal{T}(\mathbf{x}, \mathbf{y}) = \mathbf{P}(\vartheta \in \Theta_0 \mid \mathbf{x}, \mathbf{y}) &= \frac{1}{f(\mathbf{x}, \mathbf{y})} \left((1 - \pi) \int_R f(\mathbf{x} \mid t) f(\mathbf{y} \mid t) g_1(t) dt + \right. \\ &\left. + \pi \int_R \left[\int_{\Theta_0} f(\mathbf{x} \mid t) f(\mathbf{y} \mid t + \theta) g_2(\theta \mid t) d\theta \right] g_1(t) dt \right). \end{aligned} \quad (2.11)$$

В частности, если проверяется гипотеза $H_0 : \theta = 0$, то

$$\mathcal{T}(\mathbf{x}, \mathbf{y}) = \frac{1 - \pi}{f(\mathbf{x}, \mathbf{y})} \int_R f(\mathbf{x} \mid t) f(\mathbf{y} \mid t) g_1(t) dt.$$

Гипотеза H_0 принимается, если $\mathcal{T}(\mathbf{x}, \mathbf{y}) \geq C$, где константа C выбирается в зависимости от целей исследования или из условия на d-апостериорную вероятность ошибки (d-риск) I-ого рода, или соответственно d-риск II-ого рода:

$$\mathcal{R}_1(C) = \mathcal{R}(d_1; C) := \mathbf{P}\{\vartheta \in \Theta_0 \mid \mathcal{T}(\mathbf{X}, \mathbf{Y}) \leq C\} \leq \beta_1,$$

$$\mathcal{R}_0(C) = \mathcal{R}(d_0; C) := \mathbf{P}\{\vartheta \notin \Theta_0 \mid \mathcal{T}(\mathbf{X}, \mathbf{Y}) > C\} \leq \beta_0,$$

здесь d_j — решение в пользу гипотезы H_j , $j = 0, 1$.

Если существует константа C , для которой d-риск I-ого рода $\mathcal{R}_1(C) = \beta_1$, то критерий с такой константой будет иметь наименьшую d-апостериорную вероятность ошибки II-ого рода $\mathcal{R}_0(C)$ и, наоборот. В противном случае, как обычно, требуется дополнительная рандомизация; см. [21].

2.3.1 Модель N-N

Рассмотрим вероятностную модель, в которой экспериментальные значения наблюдений экспрессии каждого гена в контрольной группе (значения X -ов) распределены по нормальному закону $\mathcal{N}(t, \sigma_x^2)$ со средним, равным истинному значению экспрессии данного гена, и фиксированной дисперсией, связанной как с применяемым средством анализа проб, так и с изменчивостью экспрессии гена от пациента к пациенту. Распределение наблюдений экспрессии этого же гена в группе заболевших (значения Y -ов) будем также предполагать нормальным $\mathcal{N}(t + \theta, \sigma_y^2)$ с параметром $\theta = 0$, если экспрессия гена не связана с заболеванием, и $\theta < 0$ или $\theta > 0$, если экспрессия гена для пациентов с заболеванием приводит к понижению или, соответственно, к повышению экспрессии гена.

Пусть значения параметров вероятностной модели наблюдений экспрессии t_1, \dots, t_k всех генов в контрольной группе есть реализации независимых случайных величин ζ_1, \dots, ζ_k , распределённых одинаково по нормальному $\mathcal{N}(\mu_0, \tau_0^2)$ закону со средним μ_0 и дисперсией τ_0^2 . Значения параметра сдвига $\theta_1, \dots, \theta_k$ также есть реализации независимых одинаково распределённых случайных величин $\vartheta_1, \dots, \vartheta_k$. Распределение случайной величины ϑ есть смесь трех вероятностных законов. Для генов, экспрессия которых не связана с заболеванием — доля таких генов равна $(1 - \pi_1 - \pi_2)$, распределение ϑ сосредоточено в нуле. Для генов, экспрессия которых понижается с заболеванием — доля таких генов равна π_1 , параметр сдвига ϑ имеет нормальное $\mathcal{N}(-\mu_1, \tau_1^2)$ распределение и соответственно для генов, экспрессия которых повышается с заболеванием — доля таких генов равна π_2 , параметр ϑ имеет нормальное $\mathcal{N}(\mu_2, \tau_2^2)$ распределение, $\mu_1, \mu_2 > 0$.

Как указано выше, оптимальное решение должно быть основано на апостериорной вероятности справедливости нулевой гипотезы (2.11). Данная вероятностная модель содержит десять мешающих параметров $(\mu_0, \tau_0^2, \pi_1, \mu_1, \tau_1^2, \pi_2, \mu_2, \tau_2^2, \sigma_x^2, \sigma_y^2)$, значения которых неизвестны и должны быть оценены по выборочным данным. Полученный эмпирический аналог статистики (2.11), очевидно, не будет монотонной функцией от статистики Стьюдента. Более того, как показывают расчёты, критерий, основанный на разности выборочных средних, не может претендовать на оптимальность. В действительности область принятия нулевой гипотезы оптимального критерия может быть выражена через специфические линейные комбинации выборочных средних.

Предположим, что мешающие параметры модели полностью известны. В этом случае достаточная статистика равна вектору выборочных средних (\bar{X}, \bar{Y}) . Тестовая статистика критерия, минимизирующего d-риск II-ого рода, при проверке односторонней гипотезы $H_1 : \theta > 0$ имеет вид (2.11) с $\Theta_0 = (0, \infty)$ и плотностями f, g_1, g_2 , соответствующими описанной здесь нормальной модели. Численные расчёты показывают, что тестовая статистика оптимального критерия не является однозначной функцией разности $\bar{X} - \bar{Y}$, поскольку различные значения вектора (\bar{X}, \bar{Y}) , имеющие совпадающие значения разности, дают различные значения апостериорной вероятности спра-

ведливости гипотезы H_0 .

С помощью перебора различных значений мешающих параметров, было установлено, что тестовая статистика зависит только от двух параметров σ_x^2 и τ_0^2 . Путём дальнейшего перебора значений параметров модели и объёмов наблюдений n, m удалось установить следующий «эмпирический факт».

Предложение 2.1. 1. При проверке гипотезы $H_0 : \theta \leq 0$ против альтернативы $H_1 : \theta > 0$ тестовая статистики (2.11), как функция от выборочных средних \bar{X}, \bar{Y} зависит от них только через их линейную комбинацию

$$S = \bar{X} - \left(1 + \frac{\sigma_x^2}{n\tau_0^2}\right) \bar{Y}.$$

Следовательно, оптимальный критерий отвергает гипотезу H_0 , если $S \leq C$.

2. Аналогичный факт имеет место и при проверке гипотезы $H_0 : \theta = 0$ при двусторонней альтернативе $H_1 : \theta \neq 0$: оптимальный критерий принимает H_0 если $c_1 \leq S \leq c_2$.

Направление дальнейших исследований

- Решение вопроса о бесконечности среднего значения момента остановки универсальной процедуры относительно безусловного распределения наблюдений.
- Обобщение результатов о моменте остановки универсальной процедуры на некоторый общий класс вероятностных моделей.
- Определение параметров усечённой универсальной процедуры, при которых она становится d -гарантийной.
- Уточнение асимптотических формул для необходимого объёма выборки в рамках регулярных вероятностных моделей.
- Вывод асимптотических формул для необходимого объёма выборки в рамках нерегулярных вероятностных моделей.
- Решение проблема минимизации среднего объёма наблюдений при различении двусторонних и многих гипотез.
- Отыскание минимаксной процедуры в задаче различения более двух гипотез.
- Теоретическое доказательство предложения о виде тестовой статистики оптимального правила в задаче множественного сравнения.

Заключение

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

Изучены возможности применения, а также свойства усечённой универсальной последовательной процедуры. Численными методами исследован вопрос о её гарантийности и эффективности 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.

Список иллюстраций

1.1	Оценка функции надёжности $\bar{F}(n)$ момента остановки ν_{un} . . .	69
1.2	Траектория S_n в границах ПКОВ и ПКСВ	71
1.3	Асимптотические и точные границы случайного блуждания ν_{un} ; В–В модель	74
1.4	Ядерная оценка априорной плотности	90
2.1	Изменение d-рисков в зависимости от критической константы .	113

Список таблиц

1.1	Сравнение асимптотических формул НОВ при малых β_0, β_1 ; N–N модель	45
1.2	Точность асимптотики n_β для G–E модели	51
1.3	Точность асимптотики n_β для B–B модели	52
1.4	Точность асимптотики n_τ для G–E модели	57
1.5	Точность асимптотики n_τ для B–B модели	59
1.6	Нижние границы для стандартных отклонений оценок параметров B–B модели	84
2.1	Характеристики последовательных процедур; N–N модель . . .	96
2.2	Характеристики последовательных процедур; N–N модель . . .	97
2.3	Сравнение трёх схем выбора на данных реального примера . . .	99
2.4	Характеристики последовательных процедур; G–E модель . . .	101
2.5	Характеристики последовательных процедур; B–B модель	103
2.6	Данные по экспрессии генов из монографии Б. Эфрона	109
2.7	d-Риски и число решений d_1 критерия $ T > c$ при различении гипотез $H_0 : \theta = 0, H_1 : \theta \neq 0$; модель с 3 параметрами	112
2.8	d-Риски и число решений d_1 критерия $T > c$ при различении гипотез $H_0 : \theta \leq 0, H_1 : \theta > 0$; модель с 3 параметрами	115
2.9	Надёжность и число решений d_\pm симметричного критерия при различении гипотез $H_- : \theta < 0, H_0 : \theta = 0, H_+ : \theta > 0$; модель с 3 параметрами	117
2.10	d-Риски и число решений d_1 критерия $ T > c$ при различении гипотез $H_0 : \theta = 0, H_1 : \theta \neq 0$; модель с 6 параметрами	119
2.11	Надёжность и число решений d_\pm несимметричного критерия при различении гипотез $H_- : \theta < 0, H_0 : \theta = 0, H_+ : \theta > 0$; модель с 6 параметрами	120

Список литературы

- [1] Беляев Ю. К. *Вероятностные методы выборочного контроля*/ Ю. К. Беляев. – М.: Наука, 1975. – 408 с.
- [2] Бернштейн С. Н. О «доверительных» вероятностях Фишера// *Известия Академии Наук СССР, серия матем.* – 1941, т. **5**. – с. 85–94.
- [3] Большев Л. Н. *Таблицы математической статистики*/ Л. Н. Большев, Н. В. Смирнов – М.: Наука, 1983. – 416 с.
- [4] Вайткус П. Исследования по математической статистике и её приложениям// *Литовский Математический Сборник* – 1980, т. **XX**, № 3. – с. 117–128.
- [5] Вальд А. *Последовательный анализ*/ А. Вальд. – М.: Физматгиз, 1960. – 328 с.
- [6] Володин И. Н. Оптимальный объём выборки в процедурах статистического вывода// *Известия ВУЗов. Математика* – 1978, № 12. – с. 33–45.
- [7] Володин И. Н. Гарантийные процедуры статистического вывода (определение объёма выборки)// *Исследования по Прикладной Математике и Информатике* – 1984, № 10. – с. 13–53.
- [8] 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.
- [9] Володин И. Н., Новиков Ан. А. Статистические оценки с асимптотически минимальным d -риском// *Теория Вероятностей и ее Применения* – 1993, т. **38**, № 1, с. 20–32.
- [10] Володин И. Н., Новиков Ан. А. Локальная асимптотическая эффективность последовательного критерия отношения вероятностей при гарантийном различении сложных гипотез// *Теория Вероятностей и ее Применения* – 1998, т. **43**, № 2, с. 209–225.

- [11] Володин И. Н., Новиков Ан. А., Тес-Санчре М. Ж. Асимптотика необходимого объёма выборки для локально асимптотически нормальных экспериментов// *Исследования по Прикладной Математике и Информатике* – 2001, № 23. – с. 45–54.
- [12] Володин И. Н., Новиков Ан. А., Симушкин С. В. Гарантийный контроль качества: апостериорный подход// *Обзорные Прикладной и Промышленной Математики* – 1994, т. 1, № 2. – с. 1–32.
- [13] Володин И. Н., Симушкин С. В. О d-апостериорном подходе к проблеме статистического вывода// 3-я Вильнюсская Международная Конференция по Теории Вероятностей и Математической Статистике, Тезисы докладов. – 1981, Vol. 1. – с. 100–101.
- [14] Володин И. Н., Симушкин С. В. Статистический вывод с минимальным d-риском// *Исследования по Прикладной Математике и Информатике* – 1984, № 11, ч. 2. – с. 25–39.
- [15] Володин И. Н., Симушкин С. В. Несмещённость и байесовость// *Известия ВУЗов. Математика* – 1987, № 1. – с. 3–7.
- [16] Володин И. Н., Симушкин С. В. Доверительное оценивание в d-апостериорном подходе// *Теория Вероятностей и ее Применения* – 1990, т. 35, № 2. – с. 242–254.
- [17] Закс Ш. *Теория статистических выводов.* (Перевод на русский язык)/ Ш. Закс. – М.: «Мир», 1975. – 776 с.
- [18] Круопис Ю. И. *Минимизация целевых функций некоторых систем контроля качества/* Ю. И. Круопис. – Ин-т Матем. Киберн. Акад. Наук Лит.ССР, Вильнюс, 1981.
- [19] Новиков Ан. А. Асимптотическая оптимальность последовательного d-гарантийного критерия// *Теория Вероятностей и ее Применения* – 1987, т. 32, № 2, с. 387–391.
- [20] Пенская М. Я. О нижних границах квадратической погрешности эмпирических байесовских оценок// *Статистические методы оценивания и*

- проверки гипотез*: Межвуз. сб. науч. тр./ – Пермь: Перм. ун-т, 1995. – с. 208–226
- [21] Симушкин С. В. Оптимальные d-гарантийные процедуры различения двух гипотез// Деп. *ВИНИТИ АН СССР*. – 1981, № 5547-81. – 47 с.
- [22] Симушкин С. В. Оптимальный объём выборки при d-гарантийном различении гипотез// *Известия ВУЗов. Математика* – 1982, № 5. – с. 47–52.
- [23] Симушкин С. В. Эмпирический d-апостериорный подход к проблеме гарантийности статистического вывода// *Известия ВУЗов. Математика* – 1983, № 11. – с. 42–58.
- [24] Феллер В. *Введение в теорию вероятностей и её приложения. Том 2*/ В. Феллер. – М.: «Мир», 1984. – 752 с.
- [25] Шеффе Г. *Дисперсионный анализ*/ Г. Шеффе. – М.: «Физматлит», 1980. – 512 с.
- [26] Ширяев А. Н. *Вероятность*/ А. Н. Ширяев. – М.: «Наука», 2004. – 576 с.
- [27] Aitken M. A. *Statistical Inference: an integrated Bayesian/likelihood approach*/ M. A. Aitkin – NY: Chapman & Yall/CRC, 2010. – 236 p.
- [28] 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.
- [29] Benjamini Y., Yekutieli D. The control of the false discovery rate in multiple testing under dependency// *The Annals of Statistics* – 2001, Vol. **29**, № 4. – p. 1165–1188.
- [30] Bogdan M., Chakrabarti A., Frommlet F., Ghosh J.K. Asymptotic bayes-optimality under sparsity of some multiple testing procedures// *The Annals of Statistics* – 2011, Vol. **39**, No. 3. – p. 1551–1579.
- [31] Carrol R. J., Hall P. Optimal rates of convergence for deconvolving a density// *Journal of the American Statistical Association* – 1988, Vol. **83**. – p. 1184–1186.

- [32] Dudoit S. *Multiple Testing Procedures with Applications to Genomics*, Springer Series in Statistics/ S. Dudoit, M.J. van der Laan. 2008, 589 c.
- [33] Efromovich S. Density Estimation for the Case of Supersmooth Measurement Error// *Journal of the American Statistical Association* – 1997, Vol. **92**, № 438. – p. 526–535.
- [34] Efron B. *Large-Scale Inference. Empirical Bayes methods for estimation, testing, and prediction*/ B. Efron. – Cambridge, New York: Cambridge University Press, 2010. – 321 p.
- [35] Efron B., Tibshirani R., Storey J.D., Tusher V. Empirical Bayes analysis of a microarray experiment// *Journal of the American Statistical Association* – 2001, Vol. **96**, № 456. – p. 1151–1160.
- [36] Elie Y., Martin T.W. Least squares cross-validation for the kernel deconvolution density estimator// *Comptes Rendus Academie des Sciences, Séries I* – 2002, Vol. **334**. – p. 509–513.
- [37] Fan J. On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems// *The Annals of Statistics* – 1991, Vol. **19**, № 3. – p. 1257–1272.
- [38] Genovese C., Wasserman L. Operating characteristics and extensions of the false discovery rate procedure// *Journal of the Royal Statistical Society: Series B* – 2002, Vol. **64**, № 3. – p. 499–517.
- [39] Genovese C., Wasserman, L. A stochastic process approach to false discovery control// *The Annals of Statistics* – 2004, Vol. **32**, № 3. – p. 1035–1061.
- [40] Ghosh B.K. *Sequential Tests of Statistical Hypotheses*/ B.K. Ghosh. – Addison-Wesley, Reading, 1970.
- [41] Koo Ja.-Jo., Chung H.-Y. Log-density estimation in linear inverse problems// *The Annals of Statistics* – 1998, Vol. **26**, № 1. – p. 335–362.
- [42] Lai T.L. Asymptotical optimality of invariant sequential probability ratio tests// *The Annals of Statistics* – 1981, Vol. **9**, № 2. – p. 318–333.

- [43] Meister A. Density estimation with normal measurement error with unknown variance// *Statistica Sinica* – 2006, Vol. **16**. – p. 195–211.
- [44] Neyman J. Two breakthroughs in the theory of statistical decision making// *Review de l'Inst. Intern. de Stat.* – 1962, Vol. **30**, №5. – p. 11–27.
- [45] Rao C. R./ *Advanced statistical methods in biometric research*/ C. R. Rao. – John Wiley and Sons, New York, 1952.
- [46] Robbins H. The empirical Bayes approach to statistical decision problems// *The Annals of Mathematical Statistics* – 1964, Vol. **35**, № 1, – p. 1–20.
- [47] Ryan T. P. *Sample size determination and power*/ T. P. Ryan. – Wiley Series in Probab. and Statist. John Wiley and Sons, Hoboken, NJ, 2013.– p. 374
- [48] Sarkar S. K. Some results on false discovery rate in stepwise multiple testing procedures// *The Annals of Statistics* – 2002, Vol. **30**, № 1. – p. 239–157.
- [49] Shaffer J. P. Multiple hypotheses-testing// *Ann. Rev. Psychol.* – 1995, Vol. **46**. – p. 561–584.
- [50] Sheather S. J. Density Estimation// *Statistica Sinica* – 2004, Vol. **19**, № 4. – p. 588–597.
- [51] Sherman E. D., Volodin I. N. Empirical estimate with uniformly minimal d -risk for Bernoulli trials success probability// *Math. and Statist. Models and Methods in Reliability*. Stat. Ind. Technol., Birkhäuser/Springer, New York. – 2010, p. 297–306.
- [52] Simes R. J. An Improved Bonferroni Procedure for Multiple Tests of Significance// *Biometrika*. – 1986, Vol. **73**, №. 3. – p. 751–754.
- [53] Sorič B. Statistical “Discoveries” and effect-size estimation// *Journal of the American Statistical Association* – 1989, Vol. **89**, № 406. – p. 608–610.
- [54] Stefanski L. A., Carroll R. J. Deconvoluting kernel density estimators// *Statistics* – 1990, Vol. **21**. – p. 169–184.
- [55] 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.

- [56] Storey J. D. The positive false discovery rate: a Bayesian interpretation and the q-value// *The Annals of Statistics* – 2003, Vol. **31**, № 6. – p. 2013–235.
- [57] Storey J. D. The optimal discovery procedure: a new approach to simultaneous significance testing// *Journal of the Royal Statistical Society: Series B* – 2007, Vol. **69**, Part 3. – p. 347–368.
- [58] Volodin I. N., Novikov An. A. Asymptotics of the necessary sample size in testing parametric hypothesis: d-posterior approach// *Mathematical Methods of Statistics* – 1998, Vol. **7**, № 1, – p. 111–121.
- [59] Volodin I. N., Simushkin S. V. *D*-posterior concept of *p*-value// *Mathematical Methods of Statistics* – 2004, Vol. **13**, № 1. – p. 108–121.
- [60] Zaykin D. V., Young S. S., Westfall P. H. Using the false discovery rate approach in the genetic dissection of complex traits: A response to Weller et al. (Letter to the Editor)// *Genetics*. – 2000, Vol. **154**. – p. 1917–1918.
- [61] Vaart van der A. W. *Asymptotic Statistics*/ A. W. van der Vaart. – Cambridge, New York: Cambridge University Press, 2000. – 440 p.

Публикации автора по теме диссертации

- [62] Симушкин Д. С. Сравнительный анализ по объёму выборки двух последовательных *d*-гарантийных процедур// *Обозрение Прикладной и Промышленной Математики* – 2011, т. **18**, № 1. – с. 91–94.
- [63] Симушкин Д. С. О точности эмпирических оценок *d*-апостериорного риска// *Обозрение Прикладной и Промышленной Математики* – 2013, т. **20**, № 2. – с. 153.
- [64] 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.
- [65] Simushkin D. S., Volodin I. N. FDR is the *d*-risk// Abstracts Comm. “SysPatho Workshop «Systems Biology and Medicine»”. St.Petersburg. – 2012, p. 88–89

- [66] 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
- [67] Simushkin D. S. Empirical estimation of d-risks at distinguishing one-sided hypotheses// *Lobachevskii Journal of Mathematics* – 2016, Vol. **37**, №4. – p. 509–514.
- [68] Simushkin D. S., Simushkin S. V., Volodin I. N. D-guaranteed discrimination of statistical hypotheses: review of results and unsolved problems// *Journal of Mathematical Science* – 2018, Vol. **228**, № 5, February. – p. 543–565
- [69] Simushkin D. S., Simushkin S. V., Volodin I. N. On the d-posterior approach to the multiple testing problem// (accepted for review in *J.Stat.Comp. Simul.* – 2019)
- [70] Simushkin D. S. Asymptotic of the necessary sample size in the two hypotheses discrimination problem// *Lobachevskii Journal of Mathematics* – 2020, Vol. **41**, № 2.