Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Вычислительный центр им. А. А. Дородницына РАН
Кафедра «Интеллектуальные системы»

Жариков Илья Николаевич

Статистические тесты однородности символьных последовательностей

03.03.01 — Прикладные математика и физика

Выпускная квалификационная работа бакалавра

Научный руководитель:
д.ф.-м.н. Воронцов Константин Вячеславович

Москва
2016

Содержание

1 Введение **4**

2 Базовые понятия **6**

3 Различные постановки задачи **7**
3.1 Постановка задачи о независимости 7
3.2 Постановка задачи о сравнении параметров распределения 8

4 Статистические тесты **9**
4.1 Тест Фишера 9
4.2 G-тест 10
4.3 Z-тест 10

5 Методы множественной поправки гипотез **11**
5.1 FWER и FDR 11
5.2 Метод Холма (Holm) 12
5.3 Метод Бенджамини-Хохберга (BH) 12

6 Данные **13**
6.1 Реальные данные 13
6.2 Синтетические данные 14

7 Вычислительные эксперименты и результаты **15**
7.1 Корректность тестов 15
7.1.1 Постановка эксперимента 15
7.1.2 Результаты 15
7.2 Мощность тестов 17
7.2.1 Постановка эксперимента 17
7.2.2 Результаты 17
7.3 Однородность кодограмм 19
7.3.1 Постановка эксперимента 19
7.3.2 Результаты 19
7.4 Однородность кодограмм пациентов 20
7.4.1 Постановка эксперимента 20
7.4.2 Результаты 20
7.5 Сравнение показаний различных приборов 21
7.5.1 Постановки экспериментов 21
7.5.2 Результаты 22

8 Заключение **23**

Аннотация

Рассматриваются проблемы формирования однородных обучающих выборок для классификации символьных последовательностей, получаемых в результате дискретизации непрерывного сигнала методами символьной динамики, и преобразуемых затем в векторы частот слов. Для проверки воспроизводимости измерений и возможности формирования обучающих выборок с помощью различных измерительных устройств применяются статистические тесты, толерантные к разреженности векторов частот слов. Разработанные методы применяются в информационном анализе электрокардиосигналов в задачах, возникающих при исследовании воспроизводимости ЭКГ-сигналов и метрологической проверкой пригодности приборов, а также для построения системы скрининговой диагностики заболеваний внутренних органов человека по электрокардиограмме.

Ключевые слова: *информационный анализ электрокардиосигналов, символьная последовательность, символьная динамика, критерий согласия, электрокардиограмма, вариабельность ритма сердца.*

1 Введение

Задачи сравнения символьных последовательностей возникают во многих прикладных областях: при анализе нуклеотидных и аминокислотных последовательностей [1], текстов естественного языка [2], текстов программ, дискретных сигналов. В зависимости от приложения для сравнения символьных последовательностей используются различные метрики. Метрика Хэмминга [3] позволяет распознавать лишь полное совпадение последовательностей или их фрагментов. Широкое распространение в биоинформатике получили методы, основанные на определении минимального числа операций вставки, замены и удаления символов для преобразования одной символьной последовательности в другую: метод выравнивания [4], основанный на алгоритме Нидлмана-Вунша [5], метод минимизации редакторского расстояния между последовательностями [6], где метрикой является расстояние Левенштейна [7]. Метод выравнивания имеет ограничения, присущие всем известным его модификациям [5, 8]. Он чувствителен к длине сравниваемых последовательностей, точность метода падает экспоненциально с ростом их длины [9]. Кроме того, он требует выбора системы штрафных (весовых) функций и опорной последовательности исходя из соображений, лежащих за пределами метода.

В задачах классификации символьных последовательностей часто используется векторизация — преобразование последовательности в вектор частот слов или символьных n-грамм. Например, в [10] для группы сравниваемых последовательностей строится их статистический предок — вероятностное распределение на множестве слов, из которого любой из сравниваемых векторов мог бы быть получен путём сэмплирования. Сравнение символьных последовательностей как векторов частот слов порождает меру близости последовательностей, толерантную к перестановкам относительно коротких фрагментов внутри последовательности. Для пары последовательностей этот метод соответствует применению статистики хи-квадрат Пирсона для проверки согласия двух эмпирических распределений. Критерий согласия проверяет гипотезу однородности символьных последовательностей в следующем смысле: «два вектора частот могли быть порождены случайным разбиением одной и той же последовательности на две подпоследовательности». Недостаток классического критерия согласия Пирсона в том, что распределение хи-квадрат справедливо лишь в асимптотике и плохо описывает поведение статистики в случае сильно разреженных частотных словарей [11].

В данной работе для проверки однородности символьных последовательностей предлагается использовать статистические тесты, толерантные к низким частотам слов: точный тест Фишера, G-тест и сравнение параметров биномиальных распределений с помощью Z-статистики. Применимость этих тестов к данной задаче проверяется на синтетических данных, построенных в предположении марковости: каждый следующий символ последовательности определяется двумя предыдущими символами и вероятностью перехода согласно наперед заданной матрице переходных вероятностей. Мощности тестов оцениваются на заведомо неоднородных данных нескольких типов. Проведённые эксперименты подтверждают несмещённость и состоятельность всех трёх тестов.

Анализ вариабельности сердечного ритма [12] основан на исследовании взаимосвязей между состояниями организма человека и статистическими свойствами последовательности RR-интервалов — расстояний между соседними R-пиками электрокардиограммы. Информационный анализ электрокардиосигналов [13] расширяет ВСР по двум направлениям. Во-первых, в анализ включаются не только RR-интервалы, но и амплитуды R-пиков. Во-вторых, последовательность знаков приращений интервалов и амплитуд преобразуется в символьную последовательность (кодограмму), обычно в 6-буквенном алфавите, затем в вектор частот триграмм размерности $6^3 = 216$ [13, 14]. Кодограммы здоровых людей и больных с различными патологиями составляют обучающую выборку, по которой строится алгоритм классификации для системы скрининговой диагностики «Скринфакс»[1] [13]. Диагностика заболеваний основана на диагностических эталонах — найденных по обучающей выборке сочетаниях триграмм, специфичных для конкретных заболеваний. Аналогичный подход, основанный на методах символьной динамики, применяется в [15].

Применение тестов однородности символьных последовательностей к кодограммам позволяет решать следующие задачи на стадии сбора данных и формирования обучающих выборок:

1. Чувствительность ЭКГ к изменениям состояния и движениям обследуемого затрудняет воспроизводимость результатов. Задача состоит в том, чтобы определить, является ли кодограмма однородной, в противном случае она не подходит для дальнейшего анализа.

2. Вторая задача возникает при использовании различных типов регистрирующих устройств. Различия в кодограммах не должны быть обусловлены различиями в конструкции приборов или в условиях регистрации сигналов. Для отработки соответствующего теста был проведён эксперимент по одновременной регистрации электрокардиограммы двумя приборами, Скринфакс и CardioQvark[2].

Были также проведены эксперименты по оцениванию однородности кодограмм, зарегистрированных у одного и того же испытуемого и у разных испытуемых, как для данных Скринфакс, так и для данных CardioQvark.

[1]Информационный портал диагностической системы «Скринфакс» — skrinfax.ru

[2]Кардиомонитор CardioQvark — www.cardioqvark.ru

2 Базовые понятия

Далее в работе под кодограммой будем понимать символьную последовательность, пример которой можно увидеть на рис. 1.

Введем несколько базовых определений, которые используются в данной работе.

Определение 2.1. *k-граммой $w = (w_0, w_1, \ldots, w_{k-1})$ называется слово, образованное k последовательными символами кодограммы S.*

Определение 2.2. *Частотой k-граммы w называется число вхождений рассматриваемой k-граммы в кодограмму $S = (s_n)_{n=1}^{N}$:*

$$n_w(S) = \sum_{n=1}^{N-k+1} \prod_{j=0}^{k-1} [s_{n+j} = w_j].$$

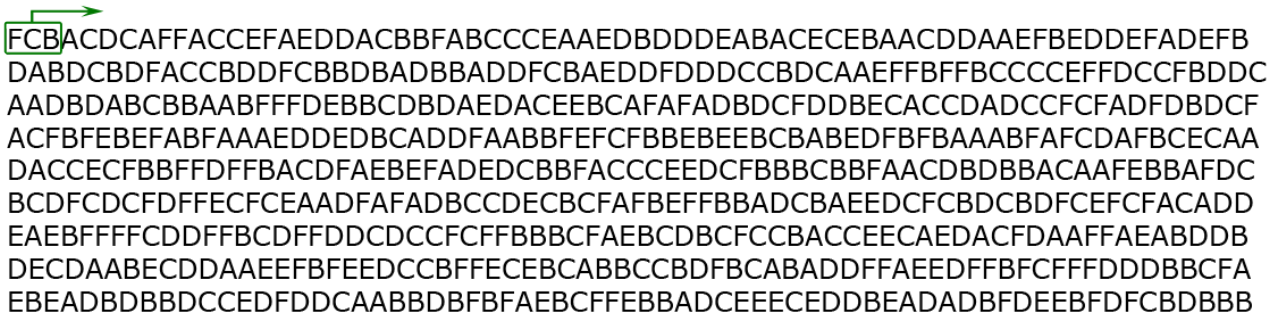


Рис. 1. Пример кодограммы, состоящей из шести различных символов.

1. DBD - 8	**10.** BDC - 6	**19.** AEB - 5	**28.** CBA - 5	**37.** EBC - 5
2. CBD - 7	**11.** BFA - 6	**20.** BAC - 5	**29.** CBB - 5	**38.** EDD - 5
3. CCE - 7	**12.** CFA - 6	**21.** BAD - 5	**30.** CCB - 5	**39.** FAC - 5
4. FAE - 7	**13.** DCC - 6	**22.** BBC - 5	**31.** CFB - 5	**40.** FAD - 5
5. FFB - 7	**14.** DDF - 6	**23.** BBF - 5	**32.** CFC - 5	**41.** FBB - 5
6. ACC - 6	**15.** DFF - 6	**24.** BCF - 5	**33.** DBB - 5	**42.** FCB - 5
7. ADB - 6	**16.** FBF - 6	**25.** BDB - 5	**34.** DCB - 5	**43.** FCF - 5
8. AED - 6	**17.** AAB - 5	**26.** BFF - 5	**35.** DCF - 5	**44.** FDD - 5
9. BBA - 6	**18.** AAE - 5	**27.** CAA - 5	**36.** DFC - 5	**45.** FFD - 5

Рис. 2. Векторное представление $n_w(S)$ кодограммы S, приведенной на рис. 1. Показаны только 45 из 216 триграмм, имеющих число вхождений $n_w(S) \geqslant 5$.

Введем в рассмотрение два множества:

$$I_S = [1, \ldots, N_S] \quad \text{и} \quad I_w = [1, \ldots, N_w]\,,$$

где множество I_S содержит номера сравниваемых символьных последовательностей, а множество I_w – номера различных k-грамм, причем нумерация отвечает за их лексикографический порядок. Очевидно, что тогда N_S — число сравниваемых последовательностей,

а N_w — число различных k-грамм, которое в общем случае равно $|\mathcal{A}|^k$, где $\mathcal{A}$ — множество символов в алфавите, из которого состоят рассматриваемые кодограммы.

В качестве признаков кодограммы S предлагается рассматривать вектор частот k-грамм:

$$\mathbf{n}(S) = \left[n_{w_1}(S), n_{w_2}(S), \ldots, n_{w_{N_w}}(S)\right]^T.$$

Для краткости обозначений через $\mathbf{n}_i$ будем обозначать $\mathbf{n}(S_i)$ — вектор частот кодограммы под номером i. Пример векторного представления кодограммы при $k = 3$ показан на рис. 2.

Таким образом, в качестве матрицы «объект-признак» рассматривается следующая матрица:

$$\mathbf{X}_{N_S \times N_w} = \begin{pmatrix} n_{11} & n_{12} & \ldots & n_{1N_w} \\ n_{21} & n_{22} & \ldots & n_{2N_w} \\ \vdots & \vdots & \ddots & \vdots \\ n_{N_S1} & n_{N_S2} & \ldots & n_{N_SN_w} \end{pmatrix} = [\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_{N_S}]^T. \tag{2.1}$$

Элемент n_{ij} матрицы $\mathbf{X}_{N_S \times N_w}$ равен частоте k-граммы w_j (k-граммы под номером j) в кодограмме S_i, то есть $n_{ij} = n_{w_j}(S_i)$, где $i \in I_S$, $j \in I_w$.

Существует несколько методов проверки символьных последовательностей на однородность, использующих разные предположения, в которых рассматриваются разные наборы векторов, соответствующие данным кодограммам. Также однородность можно интерпретировать по разному и использовать различные тесты. В связи с этим необходимо сформулировать несколько постановок задачи сравнения символьных последовательностей и обозначить статистические тесты, используемые для их решения.

3 Различные постановки задачи

3.1 Постановка задачи о независимости

Однородность символьных последовательностей эквивалентна совпадению распределений частот k-грамм в исследуемых кодограммах, что в свою очередь равносильно независимости номера k-граммы от номера кодограммы, в которой она присутствует. Для формализации этой идеи требуется ввести пару векторов, которая будет проверяться на независимость с помощью предложенных статистических тестов.

Рассматриваются два вектора:

$$\mathbf{X_S} = [i_{S1}, i_{S2}, \ldots, i_{Sl}] \quad \text{и} \quad \mathbf{X_w} = [j_{w1}, j_{w2}, \ldots, j_{wl}],$$

где $i_{Sm} \in I_S$, а $j_{wm} \in I_w$ для $m = 1, \ldots, l$. То есть первый вектор $\mathbf{X_S}$ состоит из номеров сравниваемых кодограмм, а второй вектор $\mathbf{X_w}$ – из номеров k-грамм. Причем пара (i_{Sm}, j_{wm}) означает, что k-грамма под номером j_{wm} встретилась в кодограмме под номером i_{Sm}. Очевидно, что независимость номера k-граммы от номера кодограммы эквивалентна независимости данных векторов.

Таким образом, проверяемая гипотеза выглядит следующим образом:

$$\begin{aligned} &\mathbf{H_0}: \quad \mathbf{X_S} \text{ и } \mathbf{X_w} \quad \text{независимы;} \\ &\mathbf{H_1}: \quad \mathbf{X_S} \text{ и } \mathbf{X_w} \quad \text{зависимы.} \end{aligned} \tag{3.1}$$

При проверке данной гипотезы, используются обобщенный точный тест Фишера и G-тест, которые основаны на анализе таблицы сопряженности. Таблица сопряжённости, или таблица контингентности — это средство представления совместного распределения двух переменных, предназначенное для исследования связи между ними. Структура таблицы сопряженности для векторов $\mathbf{X_S}$ и $\mathbf{X_w}$ представлена в табл. 1.

Таблица 1. Таблица $\mathbf{T}$ сопряженности размера $N_S \times N_w$.

$\mathbf{n_{+\bullet}} = \sum_{i=1}^{N_S} n_{i\bullet}$

$\mathbf{n_{\bullet+}} = \sum_{j=1}^{N_w} n_{\bullet j}$

$\mathbf{n} = \sum_{i=1}^{N_S} \sum_{j=1}^{N_w} n_{ij}$

$\mathbf{X_S}$ \ $\mathbf{X_w}$	1	2	…	$\mathbf{N_w}$	$\mathbf{n_{\bullet+}}$
1	n_{11}	n_{12}	…	n_{1N_w}	$\mathbf{n_{1+}}$
2	n_{21}	n_{22}	…	n_{2N_w}	$\mathbf{n_{2+}}$
⋮	⋮	⋮	⋱	⋮	⋮
$\mathbf{N_S}$	n_{N_S1}	n_{11}	…	$n_{N_SN_w}$	$\mathbf{n_{N_S+}}$
$\mathbf{n_{+\bullet}}$	$\mathbf{n_{+1}}$	$\mathbf{n_{+2}}$	…	$\mathbf{n_{+n_w}}$	$\mathbf{n}$

3.2 Постановка задачи о сравнении параметров распределения

При постановке данной задачи предполагается, что частота любой из k-грамм описывается биномиальным распределением, а именно:

$$n_{ij} = n_{w_j}(S_i) \sim Bin(p_{ij}, L_i), \text{ где } p_{ij} = \frac{n_{w_j}(S_i)}{L_i}, \text{ а } L_i = \sum_{j=1}^{N_w} n_{w_j}(S_i).$$

То есть p_{ij} — вероятность обнаружить k-грамму w_j в кодограмме S_i, а L_i — общее число k-грамм в рассматриваемой кодограмме.

Таким образом, для проверки однородности символьных последовательностей, в данном случае для каждой k-граммы, нужно проверить гипотезу о том, что частоты данной k-граммы в разных кодограммах пришли из одного и того же биномиального распределения. Запишем это в наших обозначениях.

Для сравнения пары символьных последовательностей рассматриваются их вектора частот k-грамм, строки матрицы (2.1):

$$\mathbf{X_1} = \mathbf{n}_i = \left[n_{w_1}(S_i), \ldots, n_{w_{N_w}}(S_i)\right]^T \quad \text{и} \quad \mathbf{X_2} = \mathbf{n}_j = \left[n_{w_1}(S_j), \ldots, n_{w_{N_w}}(S_j)\right]^T,$$

где $i \neq j$ и $i, j \in I_S$ — номера кодограмм.
И для всех $m = 1, \ldots, N_w$ проверяется гипотеза о равенстве параметров распределений

величин n_{im} и n_{jm}:

$$\begin{aligned} &\mathbf{H_0}: \quad p_{im} = p_{jm}; \\ &\mathbf{H_1}: \quad p_{im} \neq p_{jm}. \end{aligned} \tag{3.2}$$

Для проверки данной гипотезы используется статистический тест, основанный на вычислении Z-статистики. Далее под Z-тестом будем понимать именно этот тест.

4 Статистические тесты

4.1 Тест Фишера

Одним из способов проверки нулевой гипотезы (3.1) является точный тест Фишера [16]. Рассматриваются всевозможные таблицы сопряженности $\{\mathbf{T}^r\}_{r=1}^N$ (их элементы обозначим через t_{ij}^r), удовлетворяющие следующим свойствам:

$$n_{i+} = \sum_{j=1}^{N_w} t_{ij}^r, \quad \text{для любого } i \in I_S;$$

$$n_{+j} = \sum_{i=1}^{N_S} t_{ij}^r, \quad \text{для любого } j \in I_w.$$

Для заданной таблицы сопряженности $\mathbf{T}$, см. табл. 1, которая очевидным образом принадлежит множеству $\{\mathbf{T}^r\}_{r=1}^N$, вычисляется значение величины P:

$$P = \frac{\prod_{i=1}^{N_S} n_{i+}! \cdot \prod_{j=1}^{N_w} n_{+j}!}{n! \cdot \prod_{i=1}^{N_S} \prod_{j=1}^{N_w} n_{ij}!}, \tag{4.1}$$

где P — это вероятность получить данную таблицу сопряженности из всех таблиц из введенного ранее множества $\{\mathbf{T}^r\}_{r=1}^N$.

Далее для каждой из исследуемых таблиц с фиксированными суммами по строкам и столбцам вычисляются вероятности P_r, $r = 1, \ldots, N$ по формуле (4.1). Рассматривается следующее множество $\mathcal{B} = \left\{ r \;\middle|\; P_r \leqslant P,\ r = 1, \ldots, N \right\}$, состоящее из номеров таблиц, для которых вычисленные вероятности меньше заданной P, и вычисляется p-value — достигаемый уровень значимости согласно следующей формуле:

$$\text{p-value} = \sum_{m \in \mathcal{B}} P_m.$$

Критерий. Если p-value меньше уровня значимости, то гипотеза о независимости векторов $\mathbf{X_S}$ и $\mathbf{X_w}$ отвергается, что эквивалентно отвержению гипотезы об однородности рассматриваемых кодограмм.

$$\boxed{\text{p-value} < \alpha \quad \Rightarrow \quad \mathbf{H_0} \text{ отвергается.}}$$

4.2 G-тест

Еще одним способом проверки нулевой гипотезы (3.1) является G-тест. По заданной таблице сопряженности $\mathbf{T}$, см. табл. 1, вычисляется значение статистики:

$$G^2\left(\mathbf{X_S}, \mathbf{X_w}\right) = 2 \cdot \sum_{j=1}^{N_w} \sum_{i=1}^{N_S} n_{ij} \ln \frac{n_{ij} n}{n_{i+} n_{+j}} \quad \text{для всех } n_{ij} \neq 0. \tag{4.2}$$

В условиях справедливости нулевой гипотезы статистика имеет распределение хи-квадрат с $(N_S - 1)(N_w - 1)$ степенями свободы. В связи с этим можно сформулировать условия, при которых нулевая гипотеза (3.1) будет отвергнута.

Критерий. Если вычисленное значение статистики $G^2\left(\mathbf{X_S}, \mathbf{X_w}\right)$ больше, чем $(1-\alpha)$-квантиль распределения $\chi^2_{(N_S-1)(N_w-1)}$, то гипотеза о независимости случайных векторов $\mathbf{X_S}$ и $\mathbf{X_w}$ отклоняется на уровне значимости α, что эквивалентно отвержению гипотезы об однородности рассматриваемых кодограмм.

$$\boxed{G^2\left(\mathbf{X_S}, \mathbf{X_w}\right) \geqslant \chi^2_{1-\alpha,(N_S-1)(N_w-1)} \quad \Rightarrow \quad \mathbf{H_0} \text{ отвергается.}}$$

4.3 Z-тест

Данный тест используется для проверки нулевой гипотезы (3.2). При нормальной аппроксимации биномиального распределения статистикой для проверки нулевой гипотезы для каждой k-граммы вычисляется величина [17], называемая Z-статистикой:

$$Z_m = \frac{\frac{n_{im}}{L_i} + \frac{1}{2L_i} - \frac{n_{jm}}{L_j} - \frac{1}{2L_j}}{\sqrt{\frac{n_{im}+n_{jm}}{L_i+L_j} \cdot \frac{L_i+L_j-n_{im}-n_{jm}}{L_i+L_j} \cdot \left(\frac{1}{L_i} + \frac{1}{L_j}\right)}}, \tag{4.3}$$

где m — номер рассматриваемой k-граммы, для которой проверяется гипотеза о равенстве параметров распределения; n_{im}, n_{jm} — частоты данной k-граммы в кодограммах S_i и S_j соответственно (в кодограммах под номерами i и j), элементы матрицы $\mathbf{X}_{N_S \times N_w}$, см. (2.1); L_i и L_j — общее число k-грамм в рассматриваемых кодограммах.

Критерий. Обозначим через U_β — β-квантиль распределения $N(0,1)$.

Если $|Z_m| \geqslant U_{1-\frac{\alpha}{2}}$, то нулевая гипотеза (3.2) отклоняется на уровне значимости α. Данная гипотеза проверяется описанным выше способом для каждой k-граммы и в случае, когда доля k-грамм, для которых она была отвергнута, больше уровня значимости α, гипотеза об однородности рассматриваемых кодограмм отклоняется.

$$\boxed{\frac{\sum_{m=1}^{N_w} \left[|Z_m| \geqslant U_{1-\frac{\alpha}{2}}\right]}{N_w} > \alpha \quad \Rightarrow \quad \mathbf{H_0} \text{ отвергается.}}$$

5 Методы множественной поправки гипотез

В данной работе гипотезы, обозначенные в разделе 3, проверяются для многих пар кодограмм. В связи с этим необходимо использовать различные поправки на множественность тестирования, описание которых приводится в данном разделе.

5.1 FWER и FDR

Введем необходимые величины, значения которых мы будем контролировать с помощью изложенных ниже методов множественной поправки гипотез.

Пусть $H_1, H_2, \ldots, H_m$ — семейство проверяемых гипотез; M_0 — множество индексов верных гипотез; $m_0 = |M_0|$ — число верных гипотез; R — число отвергаемых гипотез; V — число ошибок первого рода. Эти и остальные обозначения приведены в табл. 2.

Таблица 2. Основные обозначения.

	Число верных H_i	Число неверных H_i	**Всего**
Число принятых H_i	U	T	$m - R$
Число отвергнутых H_i	V	S	R
Всего	m_0	$m - m_0$	m

Определение 5.1. *Групповой вероятностью ошибки первого рода (familywise error rate) называется величина:*

$$FWER = P\left(V > 0\right).$$

Контроль над групповой вероятностью ошибки на уровне α означает:

$$\mathrm{FWER} \leqslant \alpha \tag{5.1}$$

Для этих целей в данной работе используется метод Холма, см. раздел 5.2.

Определение 5.2. *Ожидаемой долей ложных отклонений гипотез (false discovery rate) называется величина:*

$$FDR = \mathbb{E}\left(\frac{V}{\max\left(R, 1\right)}\right).$$

Контроль над ожидаемой долей ложных отклонений гипотез на уровне α означает:

$$\mathrm{FDR} \leqslant \alpha \tag{5.2}$$

Для контроля за FDR в работе используется метод Бенджамини-Хохберга, см. раздел 5.3.

Замечание 5.1. Для любой процедуры множественной проверки гипотез $\mathrm{FDR} \leqslant \mathrm{FWER}$.

5.2 Метод Холма (Holm)

Пусть $p_{(1)} \leqslant p_{(2)} \leqslant \ldots \leqslant p_{(m)}$ — достигаемые уровни значимости, упорядоченные по возрастанию. Процедура Холма устроена следующим образом:

Шаг 1. Если $p_{(1)} \geqslant \dfrac{\alpha}{m}$, то мы принимаем гипотезы $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$ и останавливаемся. В противном случае, отвергаем гипотезу $H_{(1)}$ и продолжаем проверку остальных гипотез на уровне значимости $\dfrac{\alpha}{m-1}$.

Шаг 2. Если $p_{(2)} \geqslant \dfrac{\alpha}{m-1}$, то мы принимаем гипотезы $H_{(2)}, H_{(3)}, \ldots, H_{(m)}$ и останавливаемся. В противном случае, отвергаем гипотезу $H_{(2)}$ и продолжаем проверку остальных гипотез на уровне значимости $\dfrac{\alpha}{m-2}$.

Шаг 3 – Шаг m – 1. Аналогично рассмотренным выше шагам.

Шаг m. Если $p_{(m)} \geqslant \alpha$, то мы принимаем гипотезу $H_{(m)}$. В противном случае, отвергаем гипотезу.

Таким образом, метод Холма — нисходящая процедура со следующими уровнями значимости:

$$\alpha_1 = \frac{\alpha}{m}, \;\; \alpha_2 = \frac{\alpha}{m-1}, \;\; \ldots, \;\; \alpha_i = \frac{\alpha}{m-i+1}, \;\; \ldots, \;\; \alpha_m = \alpha,$$

Процедура контролирует FWER на уровне значимости α при любом характере зависимостей между достигаемыми уровнями значимости.

5.3 Метод Бенджамини-Хохберга (BH)

В данном разделе используются те же обозначения, что и в разделах 5.1 и 5.2. Процедура Бенджамини-Хохберга устроена следующим образом:

Шаг 1. Если $p_{(m)} \leqslant \alpha$, то мы отвергаем гипотезы $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$ и останавливаемся. В противном случае, принимаем гипотезу $H_{(m)}$ и продолжаем проверку остальных гипотез на уровне значимости $\dfrac{(m-1)\,\alpha}{m}$.

Шаг 2. Если $p_{(m-1)} \leqslant \dfrac{(m-1)\,\alpha}{m}$, то мы отвергаем гипотезы $H_{(1)}, H_{(2)}, \ldots, H_{(m-1)}$ и останавливаемся. В противном случае, принимаем гипотезу $H_{(m-1)}$ и продолжаем проверку остальных гипотез на уровне значимости $\dfrac{(m-2)\,\alpha}{m}$.

Шаг 3 – Шаг m – 1. Аналогично рассмотренным выше шагам.

Шаг m. Если $p_{(1)} \leqslant \dfrac{\alpha}{m}$, то мы отвергаем гипотезу $H_{(1)}$. В противном случае, принимаем гипотезу.

Таким образом, процедура Бенджамини-Хохберга является восходящей со следующими уровнями значимости:

$$\alpha_1 = \frac{\alpha}{m},\ \alpha_2 = \frac{2\alpha}{m},\ \ldots,\ \alpha_i = \frac{i\alpha}{m},\ \ldots,\ \alpha_m = \alpha,$$

Метод контролирует FDR на уровне значимости α при следующих ограничениях на статистики гипотез:

1. Статистики T_i независимы;

2. $P\left(X \in D \,|\, T_i = x\right)$ не убывает по x для $\forall i \in M_0$, где D — произвольное возрастающее множество, то есть, такое, что из $x \in D$ и $y \geqslant x$ следует, что $y \in D$.

6 Данные

6.1 Реальные данные

В данной работе предложенные тесты применяются в задачах, возникающих в информационном анализе электрокардиосигналов при исследовании воспроизводимости ЭКГ-сигналов и для метрологической проверки пригодности различных приборов.

Для начала необходимо объяснить принцип построения кодограммы, то есть символьной последовательности, по электрокардиограмме (ЭКГ), пример которой представлен на рис. 3.

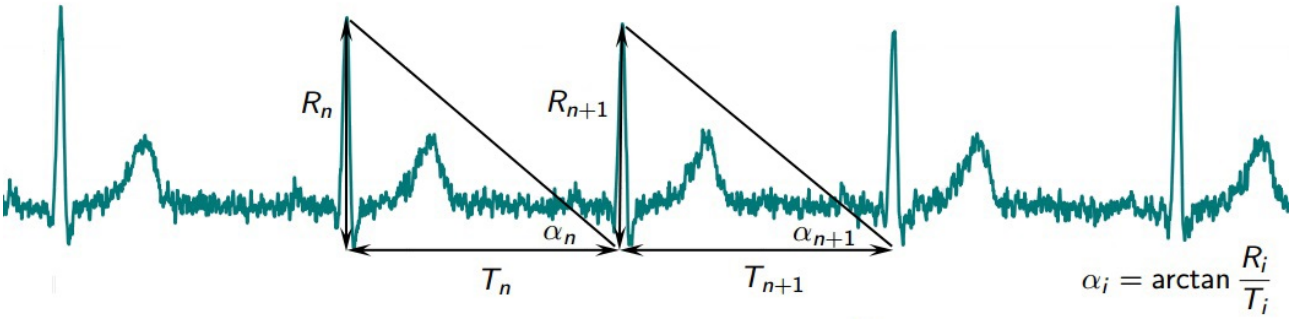


Рис. 3. Несколько кардиоциклов электрокардиограммы.

Основная идея заключается в кодировании кардиоциклов буквами из алфавита $\mathcal{A} = \{A, B, C, D, E, F\}$. Это делается согласно следующей процедуре. Каждому кардиоциклу ставится в соответствие вектор из знаков приращений трех величин: RR-интервала (T), амплитуды R-зубца (R) и угла $\alpha = \arctan \frac{R}{T}$, то есть для n-ого кардиоцикла описанный вектор выглядит следующим образом: $\left(\operatorname{sgn}\left(R_{n+1} - R_n\right), \operatorname{sgn}\left(T_{n+1} - T_n\right), \operatorname{sgn}\left(\alpha_{n+1} - \alpha_n\right)\right)$ Очевидно, что различных векторов может быть ровно 6, каждый из которых и предлагается кодировать разными буквами алфавита $\mathcal{A}$. Далее будем отождествлять понятия

электрокардиограммы и кодограммы, так как из первого однозначно получается второе, а предложенные тесты применяются именно к символьным последовательностям.

В табл. 3 представлены все наборы данных, которые используются в описанных ниже экспериментах, см. раздел 7.

Таблица 3. Используемые наборы данных

Обозначение	Количество	Описание
S	$\|S\| = 7626$	Набор кодограмм, полученных с помощью прибора Скринфакс
C	$\|C\| = 4918$	Набор кодограмм, полученных с помощью прибора CardioQvark
E	$\|E\| = 2 \cdot 23$	Набор кодограмм, полученных при одновременной записи сигнала с помощью приборов Скринфакс и CardioQvark
M	$\|M\| = 1000$	Синтетические данные (см. раздел 6.2)

6.2 Синтетические данные

Для проверки корректности статистических тестов предлагается проверять их на заведомо однородных символьных последовательностях — синтетической выборке.

Рассматривается $\mathcal{A} = \{A, B, C, D, E, F\}$ — алфавит, которому принадлежат символы из рассматриваемых кодограмм. После соединения всех символьных последовательностей множества **S**, подсчитываются величины частот встречаемости всевозможных сочетаний из трех символов алфавита $\mathcal{A}$, обозначим эти величины через p_{vb}, где $b \in \mathcal{A}$, а $v \in \mathcal{A} \times \mathcal{A}$. Затем эти величины нормируются, так чтобы

$$\sum_{b \in \mathcal{A}} p_{vb} = 1, \quad \forall v \in \mathcal{A} \times \mathcal{A}.$$

Таким образом, строится матрица переходных вероятностей, структура которой представлена в табл. 4.

Таблица 4. Матрица переходных вероятностей (размера 36×6).

	A	**B**	**C**	**D**	**E**	**F**
AA	p_{AAA}	p_{AAB}	p_{AAC}	p_{AAD}	p_{AAE}	p_{AAF}
AB	p_{ABA}	p_{ABB}	p_{ABC}	p_{ABD}	p_{ABE}	p_{ABF}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
FF	p_{FFA}	p_{FFB}	p_{FFC}	p_{FFD}	p_{FFD}	p_{FFF}

Данные генерируются согласно следующей процедуре:

Шаг 1. Первый и второй символ каждой кодограммы выбирается с одинаковой вероятностью из возможных шести.

Шаг 2. В зависимости от двух последних символов кодограммы, согласно вычисленной матрице переходных вероятностей, см. табл. 4, выбирается следующий символ из распределения, заданного соответствующей строкой данной матрицы.

Шаг 3. Повторяются действия **Шага 2** до тех пор, пока кодограмма не достигнет нужной длины.

7 Вычислительные эксперименты и результаты

Во всех последующих экспериментах в качестве признаков кодограмм использовались частоты триграмм. Различных триграмм в символьной последовательности, состоящей из 6 разных символов, может быть ровно $6^3 = 216$.

7.1 Корректность тестов

Данный эксперимент проводился для подтверждения корректности рассматриваемых тестов и их применимости к задачам, связанным с информационным анализом ЭКГ.

7.1.1 Постановка эксперимента

Данные. В данном эксперименте использовались синтетические данные $\mathbf{M}$, состоящие из 350 символов.

Эксперимент. На каждом этапе выполнялись следующие действия:

1. Выбирались случайных образом две кодограммы из множества $\mathbf{M}$.

2. Полученные кодограммы сравнивались с помощью теста Фишера, G-теста и Z-теста, причем при использовании Z-теста использовались 2 способа вычисления доли отвергнутых триграмм, см. раздел 4.3. Один из способов вычисляет долю отвергнутых триграмм только среди триграмм, у которых частоты не равны нулю хотя бы в одной из кодограмм, другой – долю триграмм среди всех 216 триграмм.

Описанные выше действия повторялись 100 раз, после чего производилась поправка на множественность тестирования для вычисленных достигаемых уровней значимости, см. раздел 5, и подсчитывалась доля пар кодограмм, для которых гипотеза об однородности была отвергнута как с учетом поправок, так и без их учета.

В эксперименте было 100 этапов, на каждом из которых доля отвергнутых пар кодограмм пересчитывалась, согласно результатам предыдущих этапов.

7.1.2 Результаты

Результаты проведенного эксперимента с использованием теста Фишера представлены на рис. 4, с использованием G-теста – на рис. 5, с использованием Z-теста – на рис. 6.

На графиках по горизонтали отображен номер этапа эксперимента, а по вертикали доля пар кодограмм, для которых гипотеза об однородности была отвергнута.

Для теста Фишера и G-теста построены три линии: результаты при использовании исходных достигаемых уровней значимости и результаты при использовании поправленных достигаемых уровней значимости с помощью методов Холма (Holm) и Бенджамини-Хохберга (BH), см. раздел 5.

Для Z-теста построены две линии: при использовании доли отвергнутых триграмм только среди триграмм, у которых частоты не равны нулю хотя бы в одной из кодограмм (Exist), и при использовании доли отвергнутых триграмм среди всех 216 триграмм (All).

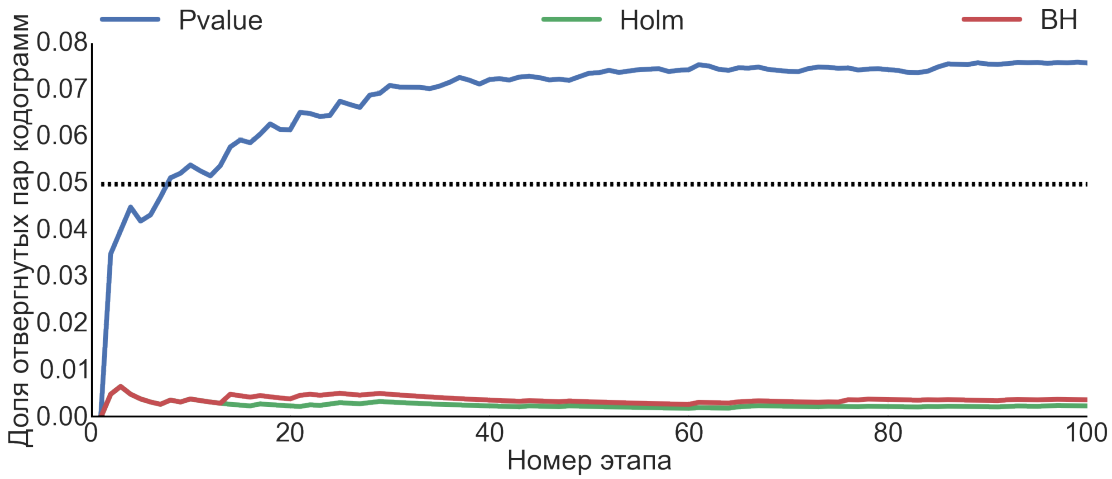


Рис. 4. Результаты, полученные с использованием теста Фишера.

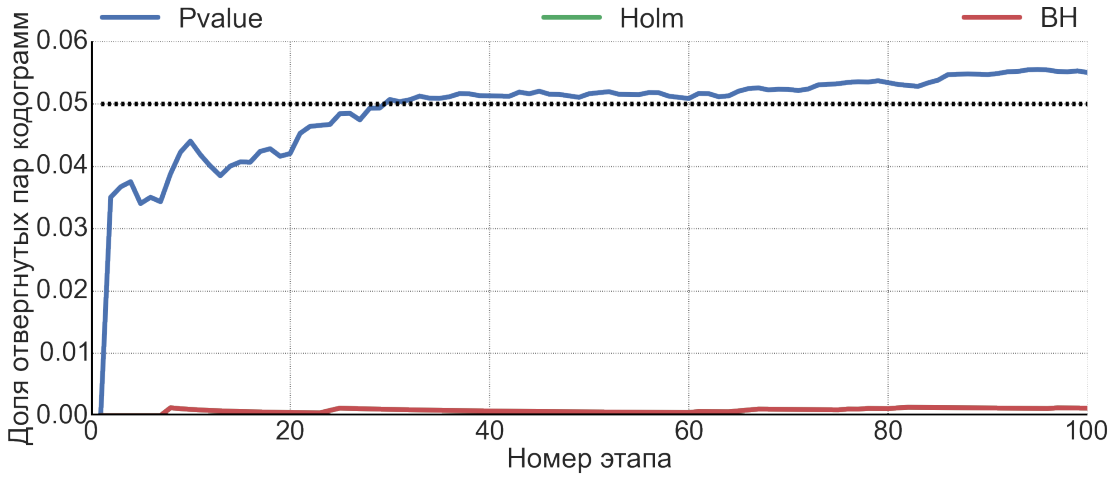


Рис. 5. Результаты, полученные с использованием G-теста.

Вывод. Все три предложенных теста являются корректными с учетом поправок на множественность тестирования.

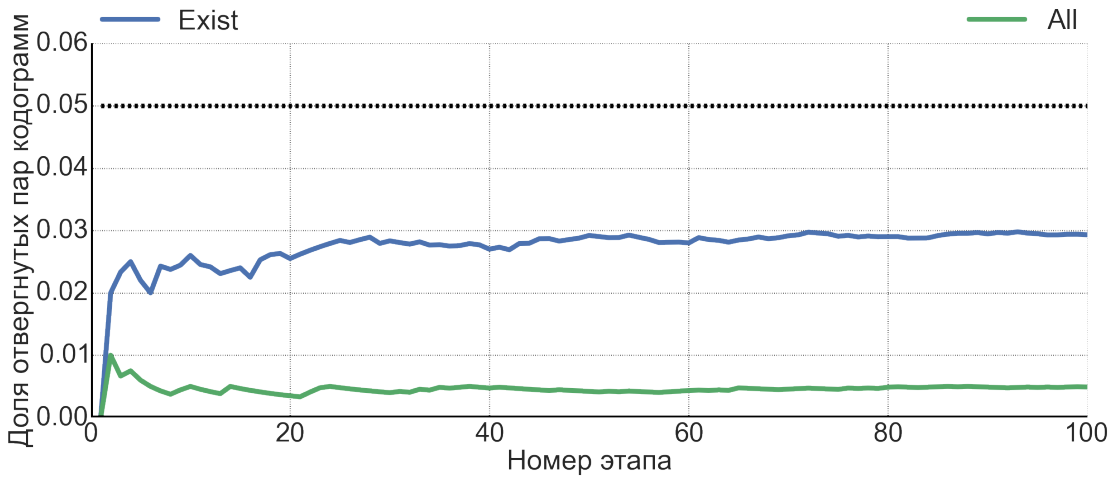


Рис. 6. Результаты, полученные с использованием Z-теста.

7.2 Мощность тестов

Данный эксперимент проводился с целью выявления наиболее мощного критерия среди трёх предложенных.

7.2.1 Постановка эксперимента

Данные. В данном эксперименте использовались реальные данные, полученные с помощью приборов Скринфакс **S** и CardioQvark **C**. Из каждого набора (**S** и **C**) случайным образом выбиралась 1000 различных кодограмм длиной 350 символов.

Эксперимент. Для заданного числа m выполнялись следующие действия:

1. Считывалась кодограмма и для нее подсчитывался вектор частот триграмм.

2. Из полученного вектора частот путем зануления m максимальных или минимальных элементов получался новый вектор частот, который сравнивался с исходным с помощью теста Фишера, G-теста и Z-теста.

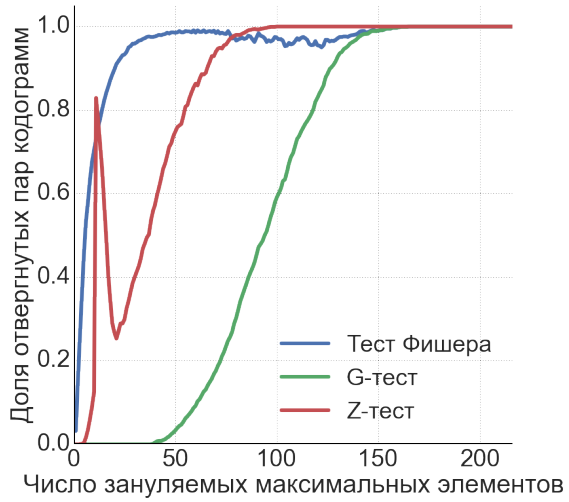
Описанные выше действия повторялись для всего набора кодограмм, после этого подсчитывалась доля пар кодограмм, для которых гипотеза об однородности была отвергнута.

Эксперимент проводился 216 раз для каждого возможного значения $m = 1, \ldots, 216$.

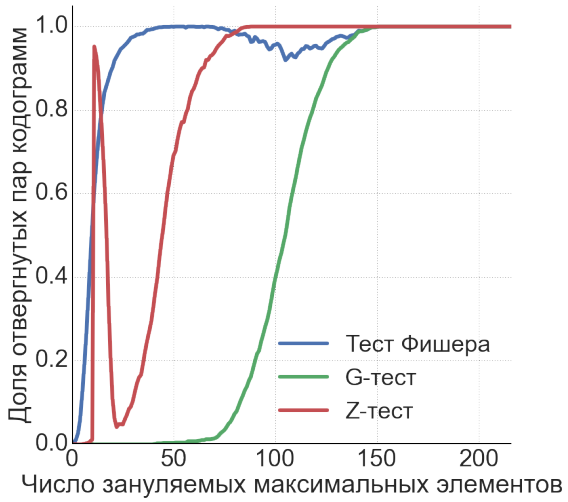
7.2.2 Результаты

Результаты проведенного эксперимента представлены в виде графиков на рис. 7. На графиках по горизонтали отображено число зануляемых элементов вектора частот триграмм, а по вертикали доля кодограмм, для которых гипотеза об однородности полученных описанным выше способом векторов частот была отвергнута.

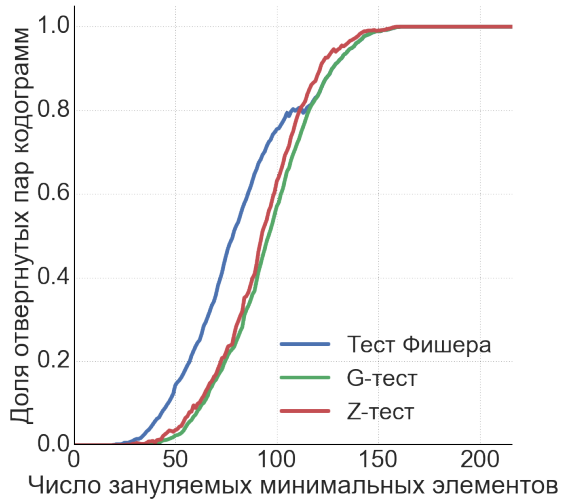
Для всех трёх тестов, изображенные линии строились по достигаемому уровню значимости без поправок на множественность тестирования, так как с учетом поправок графики качественно не изменяются.



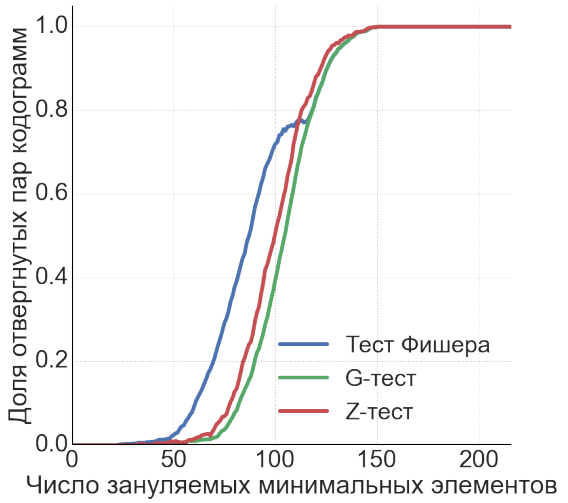
(a) Данные Скринфакс.



(b) Данные CardioQvark.



(c) Данные Скринфакс.



(d) Данные CardioQvark.

Рис. 7. Результаты тестов на данных Скринфакс и CardioQvark для двух типов неоднородных данных.

Вывод. Как видно из графиков, представленных на рис. 7, тест Фишера является наиболее мощным среди всех трёх статистических тестов на предложенных наборах неоднородных данных.

7.3 Однородность кодограмм

Данный эксперимент проводился с целью проверки однородности кодограмм в пределах одного обследования, то есть однородности кодограмм по отдельности.

7.3.1 Постановка эксперимента

Данные. В данном эксперименте использовались реальные данные, полученные с помощью приборов Скринфакс **S** и CardioQvark **C**. Из каждого набора (**S** и **C**) случайным образом выбиралась 1000 различных кодограмм длиной 350 символов.

Эксперимент. Для каждой кодограммы выполнялись следующие действия: кодограмма разбивалась на две половины, и полученная пара кодограмм сравнивалась с помощью теста Фишера, G-теста и Z-теста.

Описанные выше действия повторялись для всего набора кодограмм, после чего производилась поправка на множественность тестирования для вычисленных достигаемых уровней значимости, см. раздел 5, и подсчитывалась доля пар кодограмм, для которых гипотеза об однородности **не** была отвергнута как с учетом поправок, так и без их учета.

7.3.2 Результаты

Результаты проведенного эксперимента представлены в табл. 5. В данной таблице показана доля кодограмм, для которых гипотеза об однородности их частей **не** была отвергнута. Результаты приведены для всех трёх тестов как с учетов поправок на множественность тестирования (Holm и BH), так и без их учета (Pvalue).

Таблица 5. Доля кодограмм, для которых гипотеза об однородности внутри одного обследования **не** была отвергнута.

Данные	**Тест**	**Pvalue**	**Holm**	**BH**
Скринфакс (**S**)	Тест Фишера	0,744	1,000	0,863
	G-тест	0,973	0,998	0,994
	Z-тест	0,974	—	—
CardioQvark (**C**)	Тест Фишера	0,887	1,000	0,977
	G-тест	0,999	0,999	0,999
	Z-тест	0,997	—	—

Вывод. Как видно из табл. 5, доля неотвергнутых кодограмм значительная согласно всем рассматриваемым тестам как на данных прибора Скринфакс, так и на данных прибора CardioQvark. Отсюда можно сделать вывод об однородности кодограмм в пределах одного обследования.

7.4 Однородность кодограмм пациентов

Данный эксперимент проводился с целью выяснения различий между кодограммами одного и того же пациента и кодограммами различных пациентов.

7.4.1 Постановка эксперимента

Данные. В данном эксперименте использовались реальные данные, полученные с помощью приборов Скринфакс **S** и CardioQvark **C** и состоящие из кодограмм длиной 350 символов.

Эксперимент. На каждой итерации для каждого набора данных выполнялись следующие действия:

1. Выбиралась пара кодограмм одного и того же пациента и пара кодограмм разных пациентов. Пациенты и их кодограммы выбирались случайно.

2. Каждая пара кодограмм сравнивалась с помощью теста Фишера, G-теста и Z-теста.

Описанные выше действия повторялись 1000 раз (1000 итераций), после чего производилась поправка на множественность тестирования для вычисленных достигаемых уровней значимости, см. раздел 5, и подсчитывалась доля пар кодограмм, для которых гипотеза об однородности **не** была отвергнута как с учетом поправок, так и без их учета.

7.4.2 Результаты

Таблица 6. Доля пар кодограмм, для которых гипотеза об однородности **не** была отвергнута.

Исследовались	**Данные**	**Тест**	**Pvalue**	**Holm**	**BH**
Кодограммы одного пациента	Скринфакс (**S**)	Тест Фишера	0,105	1,000	0,109
		G-тест	0,290	0,400	0,292
		Z-тест	0,255	—	—
	CardioQvark (**C**)	Тест Фишера	0,356	1,000	0,388
		G-тест	0,760	0,903	0,830
		Z-тест	0,728	—	—
Кодограммы разных пациентов	Скринфакс (**S**)	Тест Фишера	0,004	1,000	0,004
		G-тест	0,047	0,097	0,049
		Z-тест	0,038	—	—
	CardioQvark (**C**)	Тест Фишера	0,077	1,000	0,080
		G-тест	0,357	0,630	0,391
		Z-тест	0,343	—	—

Результаты проведенного эксперимента представлены в табл. 6. В данной таблице приведены значения доли пар кодограмм, для которых гипотеза об однородности **не** была

отвергнута. Результаты приведены для всех трех тестов как с учетов поправок на множественность (Holm и BH), так и без их учета (Pvalue).

Вывод. Из табл. 6 видно, что доля неотвергнутых пар кодограмм разных пациентов существенно меньше доли неотвергнутых пар кодограмм одного пациента, то есть кодограммы разных пациентов более неоднородны, чем кодограммы одного и того же пациента. Также можно заметить, что и кодограммы одного пациента не всегда однородны.

7.5 Сравнение показаний различных приборов

Данный эксперимент проводился с целью сравнения показаний приборов Скринфакс и CardioQvark, а именно проверки однородности показаний этих приборов.

7.5.1 Постановки экспериментов

Данные. В данном эксперименте использовались реальные данные, полученные при одновременной записи с помощью приборов Скринфакс и CardioQvark **E** и состоящие из кодограмм длиной 350 символов.

Эксперимент 1. В этом эксперименте статистические тесты использовались для сравнения кодограмм одного пациента, полученных с помощью разных приборов. Количество таких пар равно 23, поэтому поправка на множественность тестирования не вычислялась.

Эксперимент 2. Данный эксперимент состоял их двух основных частей и проводился для подтверждения результатов эксперимента 1.

Часть 1. Для каждой пары синхронизированных кодограмм вычислялись разности таких величин, как RR-интервалы (T), амплитуды R-зубцов (R) и вектора частот триграмм (n). Обозначим описанные множества через $\triangle \mathrm{T}^i$, $\triangle \mathrm{R}^i$, $\triangle \mathrm{n}^i$ где $i = 1, \ldots, M$, а M — число синхронизированных пар кодограмм (в нашем случае $M = 23$). Также рассматривались множества, задающие эмпирическое распределение указанных выше величин:

$$\triangle \mathrm{T}_{all} = \bigcup_{i=1}^{M} \triangle \mathrm{T}^i, \quad \triangle \mathrm{R}_{all} = \bigcup_{i=1}^{M} \triangle \mathrm{R}^i, \quad \triangle \mathrm{n}_{all} = \bigcup_{i=1}^{M} \triangle \mathrm{n}^i.$$

После этого с помощью критерия Смирнова проверялась гипотеза об однородности разностей каждой из трёх исследуемых величин, множества сравнивались попарно.

Часть 2. Дальнейшие действия выполнялись с величиной, разности которой оказались однородными согласно результатам первой части данного эксперимента.
Для получения зашумленных данных из эмпирического распределения данной величины с помощью bootstrap'а генерировался шум и накладывался на исходные данные. Таким образом получались данные 2-х типов: зашумленные и исходные. Далее

алгоритмы классификации болезней внутренних органов человека по электрокардиограмме — случайный лес (RF), синдромный алгоритм (SA) и логистическая регрессия на главных компонентах (LR_PCA) запускались для задачи классификации здоровый-больной для 45 различных болезней. Скользящий контроль проводился 3 раза по 10 блокам. Значения параметров, максимизирующие AUC на контроле, считались оптимальными. Обучение и контроль производились как на данных одного типа, так и на данных разного типа.
Качество классификации измерялось с помощью AUC.

7.5.2 Результаты

Эксперимент 1. Из 23 пар кодограмм тестом Фишера гипотеза однородности была отвергнута для 7 пар кодограмм, G-тестом – для 1 пары кодограмм, Z-тестом – ни для одной пары кодограмм.

Вывод. Таким образом, можно сделать вывод об однородности показаний приборов, то есть данные, полученные с двух приборов, можно смешивать при формировании обучающих выборок.

Эксперимент 2. Результаты проведенного в **части 1** эксперимента представлены в табл. 7. В данной таблице показаны доли пар сравниваемых множеств разностей исследуемых величин, для которых гипотеза об однородности была отвергнута как с учетом, так и без учета поправок на множественность тестирования. Согласно результатам можно сделать вывод о том, что разности частот триграмм пришли из одного распределения, в качестве которого рассматривалось их эмпирическое распределение $\triangle \mathrm{n}_{all}$. Именно из него генерировался шум на исходные данные во второй части данного эксперимента.

Таблица 7. Доля пар кодограмм, для которых гипотеза о равенстве распределений согласно критерию Смирнова была отвергнута.

	p-value	**Holm**	**BH**
$\triangle$n	0,156	0,004	0,036
$\triangle$R	0,996	0,996	0,996
$\triangle$T	0,993	0,989	0,993

Результаты эксперимента, проведенного в **части 2** продемонстрированы в табл. 8 и табл. 9. На первом этапе обучение алгоритмов классификации производилось на зашумленных и исходных данных независимо. Скользящий контроль проводился 3 раза по 10 блокам, на каждой итерации вычислялись значения AUC на контрольной и обучающей подвыборках, затем полученные значения усреднялись. В табл. 8 приведены средние значения AUC (всего было 45 задач классификации), максимальная по модулю разница показателей качества и её среднее значение.

На втором этапе использовались алгоритмы классификации, настроенные на первом этапе как на зашумленных, так и на исходных данных. Тестирование полученных алго-

Таблица 8. Средние значения AUC на обучении и на контроле на данных одного типа.

	Обучение и контроль на данных одного типа					
	AUC на обучении			AUC на контроле		
	SA	LR_PCA	RF	SA	LR_PCA	RF
Без шума	0,86169	0,95741	1,00000	0,86065	0,93676	0,95423
С шумом	0,86220	0,95553	1,00000	0,86137	0,93508	0,94415
Максимальная разница	0,04267	0,02311	0,00000	0,04518	0,00878	0,05531
Средняя разница (без шума - с шумом)	-0,00051	0,00188	0,00000	-0,00072	0,00168	0,01008

Таблица 9. Средние значения AUC на контроле при обучении на данных разного типа.

	Обучение на исходных данных			**Обучение на зашумленных данных**		
	AUC на контроле			AUC на контроле		
	SA	LR_PCA	RF	SA	LR_PCA	RF
Без шума	0,86234	0,95531	1,00000	0,86195	0,95459	0,98233
С шумом	0,86189	0,95232	0,99370	0,86200	0,95239	0,97435
Максимальная разница	0,00394	0,01192	0,01513	0,00602	0,00942	0,04692
Средняя разница (без шума - с шумом)	0,00045	0,00299	0,00631	0,00005	0,00220	0,00798

ритмов проводилось так же на данных разного типа. Результаты представлены в табл. 9.

Вывод. Как видно из табл. 8 наложение шума не ухудшает качество классификации. Следовательно, для данных различных приборов можно применять одни и те же алгоритмы классификации. Результаты, представленные в табл. 9, позволяют говорить о том, что данные одного прибора годятся для пополнения обучающей выборки другого.

8 Заключение

В данной работе предложены статистические тесты для проверки однородности символьных последовательностей, также показаны их корректность и применимость к задачам информационного анализа электрокардиосигналов, на данных которой они тестировались. Показано, что кодограммы в пределах одного обследования, как правило, однородны. Было выяснено, что кодограммы разных пациентов более неоднородны, чем кодограммы одного и того же пациента. Однако обследования определенного человека могут также давать неоднородные кодограммы. В экспериментах с параллельной регистрацией ЭКГ двумя приборами показана однородность кодограмм, полученных с помощью систем Скринфакс и CardioQvark, и подтверждена правильность вывода путем запуска алгоритмов классификации болезней внутренних органов человека на данных прибора Скринфакс и на смоделированных данных прибора CardioQvark.

Список литературы

[1] D. Gusfield. *Algorithms on strings, trees and sequences: computer science and computational biology* // Cambridge university press, 1997. — P. 326.

[2] S. Brin, J. Davis, H. Garcia-Molina. *Copy detection mechanisms for digital documents* // ACM SIGMOD Record. — 1995. — Vol. 24, no. 2. — P. 398–409.

[3] R. W. Hamming. *Coding and information theory* // Prentice-Hall, 1986.

[4] F. Cristino et al. *ScanMatch: A novel method for comparing fixation sequences* // Behavior research methods. — 2010. — Vol. 42, no. 3. — P. 692–700.

[5] S. B. Needleman, C. D. Wunsch. *A general method applicable to the search for similarities in the amino acid sequence of two proteins* // Journal of molecular biology. — 1970. — Vol. 48, no. 3. — P. 443–453.

[6] А. М. Соколов. *Исследование ускоренного поиска близких текстовых последовательностей с помощью векторных представлений* // Кибернетика и системный анализ. — 2008. — с. 32–47.

[7] В. И. Левенштейн. *Двоичные коды с исправлением выпадений, вставок и замещений символов* // Доклады Академий Наук СССР. — 1965. — Т. 163, № 4. — с. 845–848.

[8] S. K. Zahid et al. *A novel structure of the Smith-Waterman Algorithm for efficient sequence alignment* // 2015 Third International Conference on. – IEEE. — 2015. — P. 6–9.

[9] М. Г. Садовский. *О сравнении символьных последовательностей* // Вычислительные технологии. — 2005. — Т. 10, № 3. — с. 108–116.

[10] М. Г. Садовский. *Информационно-статистический анализ нуклеотидных последовательностей* // дис. — Ин-т биофизики СО РАН. — 2004.

[11] В. Р. Целых, К. В. Воронцов. *Критерии согласия для разреженных дискретных распределений и их применение в тематическом моделировании* // Машинное обучение и анализ данных. — 2012. — Т. 1, № 4. — с. 437–447.

[12] Р. М. Баевский, Г. Г. Иванов. *Вариабельность сердечного ритма: теоретические аспекты и возможности клинического применения* // Ультразвуковая и функциональная диагностика. — 2001. — №. 3. — с. 108–127.

[13] В. М. Успенский. *Информационная функция сердца* // Клиническая медицина. — 2008. — Т. 86, № 5. — с. 4–13.

[14] V. M. Uspenskiy, K. V. Vorontsov, V. R. Tselykh et al. *Information function of the heart: discrete and fuzzy encoding of the ECG-signal for multidisease diagnostic system* // Advanced Mathematical and Computational Tools in Metrology and Testing X. — 2015. — Vol. 10. — P. 377–384.

[15] U. Parlitz, S. Berg, S. Luther et al. *Classifying cardiac biosignals using ordinal pattern statistics and symbolic dynamics* // Computers in Biology and Medicine. — 2012. — Vol. 42, no. 3. — P. 319–327.

[16] R. M. Cyrus. *A network algorithm for performing Fisher's exact test in r × c contingency tables* // Journal of the American Statistical Association. — 1983. — Vol. 78, no. 382. — P. 427–434.

[17] А. И. Кобзарь. *Прикладная математическая статистика. Для инженеров и научных работников*. — 2006. — с. 816.